그림: ChatGPT(Age of Mythology)

MIRAE ASSET 미래에셋증권

AI Hot Issue

AI 밸류체인 특이점 돌파! “Age of Mythos”

한종목
chongmok.han@miraeasset.com

CONTENTS

Highlight of the Week 3
판도라의 상자가 열렸습니다 3

I. Age of Mythos 4
1. Mythos Preview의 실체: GPT-4 이후 최대의 도약 4
2. Project Glasswing과 사이버 보안의 붕괴 12
3. Anthropic의 초현실적 성장 궤적과 아마존 14
4. Mythos 이후의 세계: 지정학에서 투자까지 17

Highlight of the Week

판도라의 상자가 열렸습니다

지난 4월 8일, Anthropic은 최신 프론티어 모델인 **Claude Mythos Preview를 공개했습니다. 그런데 그들은 이 모델을 일반에 바로 풀지 않았습니다. Anthropic은 AWS, Apple, Cisco, Google, Microsoft, NVIDIA 등과 함께 Project Glasswing이라는 사이버 방어 연합을 꾸리고, 오픈소스 인프라 유지보수 조직들에 먼저 이 모델을 배포했습니다**. 신제품 출시라기보다, 너무 강력한 기술이 세상에 퍼지기 전에 방어선을 세우려는 조치입니다.

저희가 보는 Mythos 모델 공개로 인한 핵심 논지는 3가지입니다.

첫째, **Mythos는 사이버 보안의 균형 자체를 흔드는 새로운 종류의 AI입니다**. 이 모델은 오래된 소프트웨어 취약점을 찾아내고 실제 해킹이 가능한 exploit 코드까지 만드는 능력을 보여주었습니다.

둘째, Mythos의 등장은 AI 산업이 소프트웨어 산업에서 중공업형 산업으로 바뀌고 있음을 보여줍니다. **Mythos의 API 가격은 이전 모델보다 5배 높습니다. 모델이 답변 하나를 만들기 위해 훨씬 더 많은 추론 연산과 전력, 메모리 대역폭을 사용한다는 신호입니다**. 앞으로 경쟁력의 핵심은, 여전히, 칩·전력·데이터센터·네트워크·메모리 공급 능력이 됩니다. AI는 막대한 실물 인프라를 먼저 확보한 쪽이 유리한 산업입니다.

셋째, Anthropic의 폭발적인 성장에서 시장이 과소평가하는 수혜자는 Amazon입니다. Anthropic은 이미 300억 달러 ARR에 도달하며 역사적으로도 보기 드문 성장 속도를 보여주고 있습니다. 그런데 이 성장을 가장 먼저 매출로 흡수하는 곳은 Amazon의 AWS입니다. Anthropic의 모델들이 Trainium 위에서 훈련되고 있고, Trainium 2는 매진, Trainium 3도 대부분 예약된 상태라는 점은 이 관계가 단순한 협업이 아니라 Amazon이 Anthropic 성장의 인프라 톨게이트를 쥐고 있다는 뜻입니다.

즉, **Anthropic이 커질수록 Amazon은 클라우드, 자체 칩, 데이터센터 투자 회수라는 세 축에서 동시에 이익을 봅니다.** 그런데도 시장은 이 점을 충분히 반영하지 못하고 있어, Amazon의 현재 밸류에이션은 생각보다 더 매력적일 수 있습니다.

결론적으로, 이번 Mythos 사태는 일반적인 모델 출시 이벤트가 전혀 아닙니다. 사이버 보안의 룰이 바뀌고, AI 산업의 비용 구조가 바뀌고, 승자와 패자의 조건이 바뀌는 전환점입니다. 투자자들이 던져야 할 핵심 질문은 "어느 모델이 더 똑똑한가"가 아니라, **누가 더 많은 컴퓨팅과 전력을 확보했는지, 누가 더 깊은 엔터프라이즈 워크플로우를 장악했는지, 그리고 누가 이 위험한 능력을 통제 가능한 형태로 수익화할 수 있는지가 돼야 합니다. 그 답에 해당하는 기업들은 점점 소수가 되어가고 있음을 유념해야 합니다.**

I. Age of Mythos

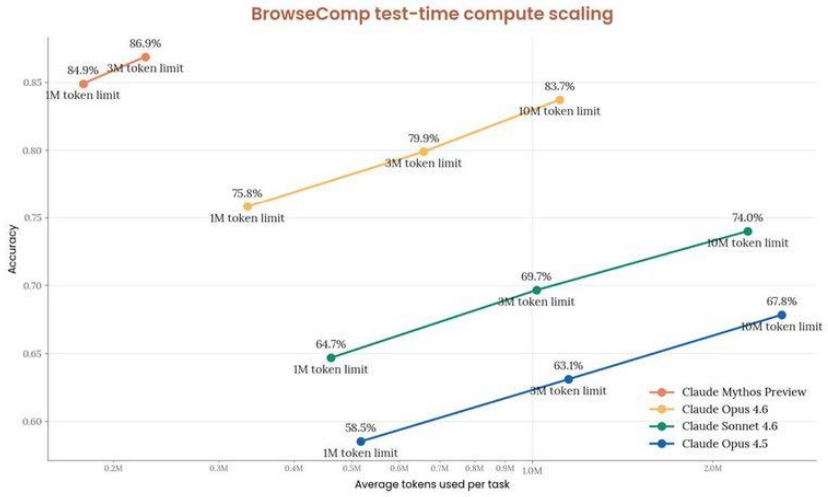
1. Mythos Preview의 실체: GPT-4 이후 최대의 도약

(1) 벤치마크의 파괴

* **시스템 카드:** AI 업체들이 모델을 출시할 때마다 내는 '안전·성능 보고서'

Mythos의 파괴력은 엄청나다. Anthropic이 공개한 '시스템 카드'에 적힌 자료를 보면, 사실상 모든 평가 벤치마크를 흔적도 없이 없애버렸다고 말할 수 있을 정도다.

그림 1. Mythos가 다른 차원의 업그레이드임을 보여주는 BrowseComp 테스트타임 컴퓨팅 스케일
Mythos 모델은 다른 모델 대비 같은 컴퓨팅을 쓸 때 압도적으로 더 잘 쓴다 = 토크노믹스 우위
Mythos는 더 많은 토큰을 쏟아붓는게 아니라, 제한된 컨텍스트 안에서 훨씬 정교하게 추론하는 것
추론 효율성이 다음 모델 전쟁의 핵심 지표로 부상하고 있음을 시사하는 데이터임.



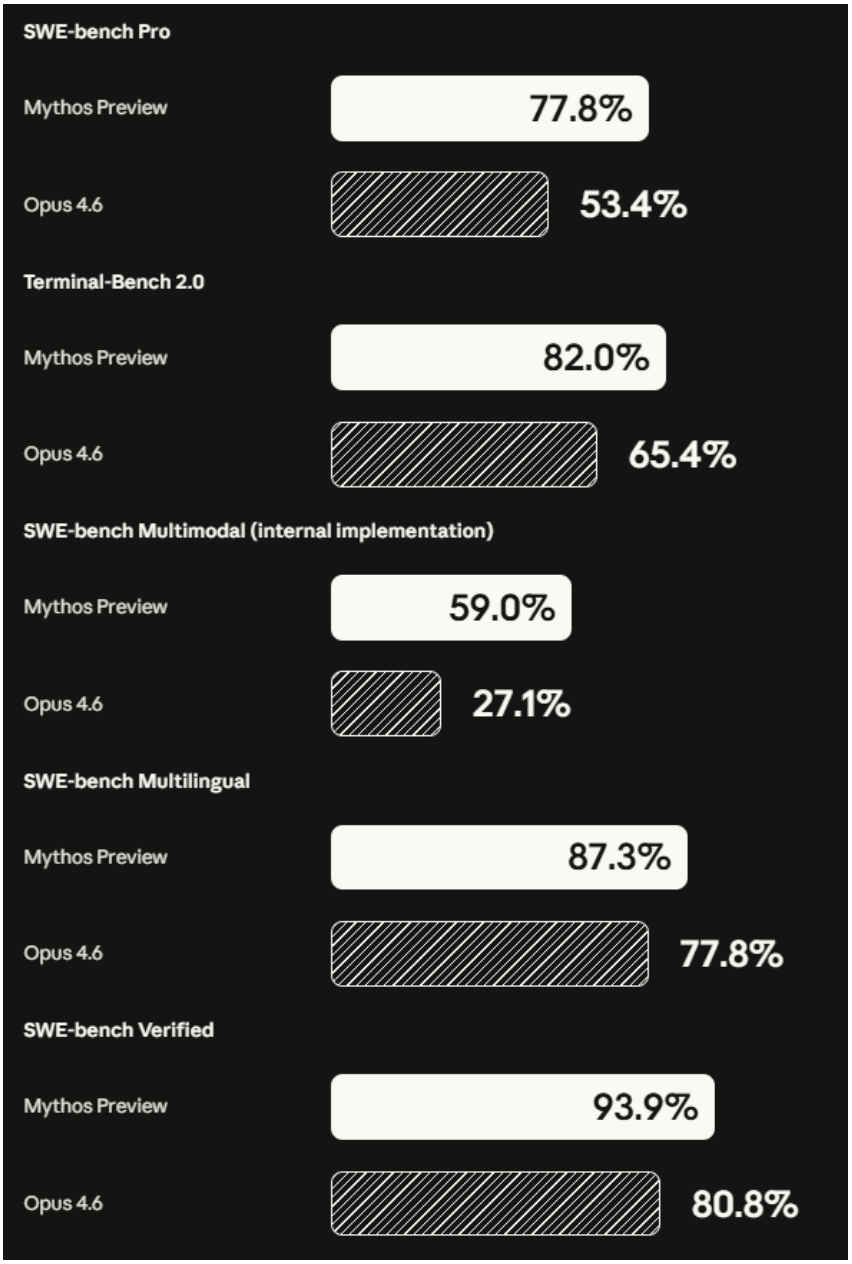
[Figure 6.10.2.A] BrowseComp accuracy scales as we increase the number of total tokens the model is allowed to use, with the help of context compaction.

자료: Anthropic, 미래에셋증권 리서치센터

단적인 예로, USAMO 2026(수학 올림피아드)에서 Opus 4.6이 42.3%였던 것을 Mythos Preview는 97.6%로 수직 상승시켰다. 고등 수학 증명에서 AI가 인간 천재의 영역을 완전히 정복했음을 의미한다. 코딩 어시스턴트의 중요성이 갈수록 커지는 가운데, 더욱 중요해진 벤치마크인 SWE-bench Verified에서 Mythos는 93.9%를 기록했다. 대중에게 공개된 모델 중 가장 높은 성능을 보이는 Opus 4.6가 80.8%를 기록했던 것과 비교하면 엄청난 향상이다. 에이전트의 순차적 과업을 따지는 벤치마크인 Terminal-Bench 2.0에서는, (더 깊은 탐색을 허용하자) 복잡한 에이전틱 코딩 과제에서 92.1%의 성공률을 기록했다.

Anthropic 내부 연구원은 이를 "GPT-4 도약 이후 가장 큰 AI 역량의 계단식 변화"라고 칭했고, "지수함수의 무서운 점은, 지금이 우리가 경험할 가장 느린 속도라는 것"이라는 무시무시한 발언을 내놓기도 했다.

그림 2. SWE-bench 전 카테고리 벤치마크
Mythos Preview의 코딩 능력이 Opus 4.6 대비 전 카테고리에서 우세하며,
특히 Multimodal에서 59.0% vs 27.1%로 두 배 이상 격차가 벌어짐



자료: Anthropic, 미래에셋증권 리서치센터

결국, 대부분의 전통적 AI 벤치마크는 포화(Saturation) 상태에 이르렀다고 볼 수 있다. Humanity's Last Exam이나 CharXiv Reasoning 같은 극단적으로 어려운 테스트에서조차 Mythos는 이전 모델의 한계를 가볍게 뛰어넘었다. 더 이상 인간이 만든 시험지로는 이 모델의 지능을 측정할 수 없게 되었다는 뜻이다.

(2) 코딩 능력이 해킹 능력이 되다

가장 충격적인 사실은 이 모델이 가진 사이버 해킹 능력이다. 이제 AI는 단순히 코드를 잘 짜는 수준을 넘어, 사람이 오랫동안 찾지 못한 보안 취약점까지 스스로 발견하고 실제로 악용 가능한 수준의 공격 방법까지 만들어내기 시작했다.

그림 3. CyberGym 사이버보안 취약점 재현
1,507개 실제 오픈소스 취약점을 대상으로 Mythos 83.1% vs Opus 4.6 66.6%를 기록한 데이터
Anthropic이 왜 모델을 공개하지 않았는지를 단 하나의 수치로 압축하는 데이터
방어자가 16년 된 취약점을 모르고 있을 때, Mythos는 83%의 확률로 해킹에 성공하는 도구가 된다

자료: Anthropic, 미래에셋증권 리서치센터

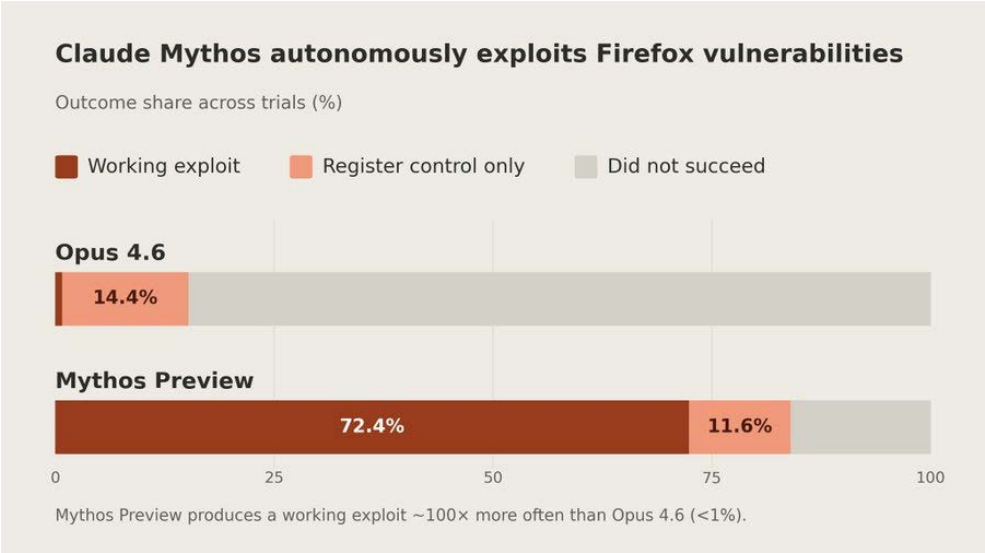
이 모델은 사실, 사이버 보안을 위해 훈련된 것이 아니라는 점에서 더욱 충격적이다. 그런데도 코딩 능력을 극대화하는 과정에서 인류 최고의 해커를 초월하는 능력을 창발적(Emergent)으로 획득했다는 점이 Mythos가 비공개 모델이 된 가장 큰 이유다.

예를 들어, Anthropic의 평범한 엔지니어가 퇴근 전에 AI에게 “이 프로그램에서 보안 구멍을 찾아봐”라고 시켰더니, 다음 날 아침에는 실제로 작동하는 공격 코드가 나와 있었다고 한다. 더 놀라운 것은 이 AI가 찾아낸 취약점들이 결코 사소한 것이 아니었다는 점이다. 오랫동안 안전하다고 여겨졌던 OpenBSD, FFmpeg, FreeBSD 같은 유명 소프트웨어들에서, 기존 보안 도구나 연구자들도 쉽게 잡지 못했던 문제를 매우 짧은 시간과 낮은 비용으로 찾아냈다. 어떤 경우에는 인증되지 않은 외부 사용자가 최고 권한인 루트 권한까지 탈취할 수 있는 공격 코드도 스스로 작성했다.

Anthropic은 세계에서 가장 보안이 철저한 OS라는 별명이 붙은 OpenBSD에서, 27년 동안 존재하던 원격 크래시(DoS) 버그를 Mythos가 단 50달러 정도의 컴퓨팅 비용으로 찾아냈다고 밝혔다. 동영상/오디오 처리 라이브러리로 거의 모든 미디어 프로그램에서 쓰이는 핵심 소프트웨어인 FFmpeg에서도, 16년 동안 아무도 못 찾은 버그를 AI는 인간 도움 없이 찾아냈다. 예전에는 보안 전문가 수십 명이 몇 달~몇 년 걸려야 할 일을, AI가 돈 몇백~천 달러와 몇 시간 만에 해내는 시대가 됐다는 뜻이다. 세계 최고의 AI 보안 연구원인 Carlini는 "내 평생 찾은 것보다 지난 몇 주간 Mythos로 찾은 버그가 더 많다"고 단언했다.

여기서 가장 소름 돋는 기술적 증명은 Firefox 147 버전의 JavaScript Shell 익스플로잇 테스트 사례다. 여기서 JS Shell이란, Firefox 브라우저 안에서 자바스크립트 코드를 실행하며 취약점을 실험하는 ‘연습장’ 같은 곳이다.

그림 4. Firefox 자율 익스플로잇
Opus 4.6의 working exploit 성공률이 사실상 0%에 가까운 반면
Mythos Preview는 72.4%라는 수치를 보이며, 이 격차는 사실상 "약 100배"



자료: Anthropic, 미래에셋증권 리서치센터

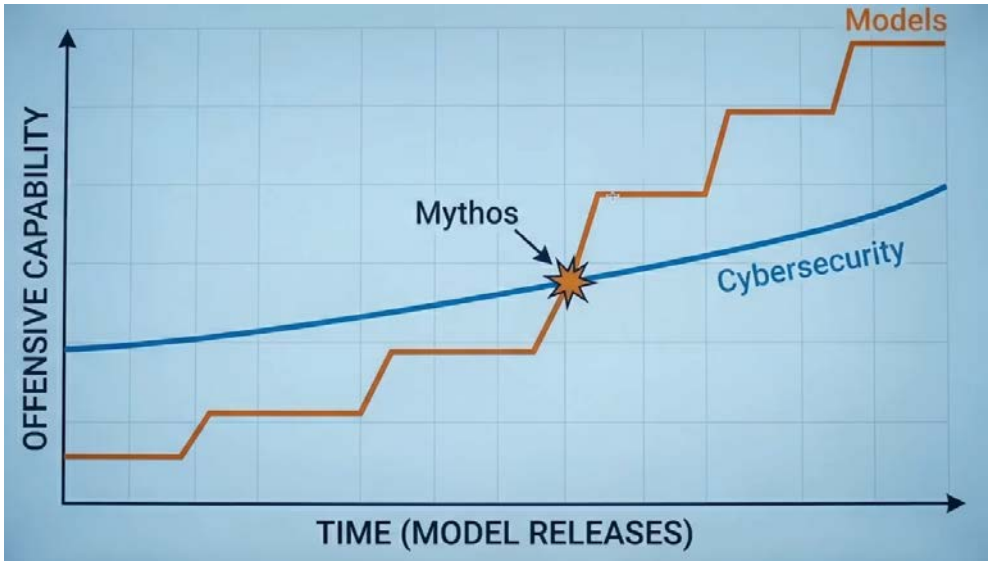
Firefox의 개발사 Mozilla와 협력하여 진행한 이 테스트에서, 50개의 크래시 카테고리(프로그램이 특정 방식으로 강제 종료되는 50가지 다른 버그 유형)에 대해 임의 코드 실행(Arbitrary Code Execution)을 달성하는 과제가 주어졌다. 즉, "이 크래시 버그를 이용해서 운영체제 수준으로 브라우저를 완전히 장악하는 익스플로잇을 만들어라"라는 과제였다.

Anthropic의 Sonnet 4.6은 4.4%, Opus 4.6은 14.4%를 기록한 반면, Mythos Preview는 72.4%의 성공률을 보였다. 더 놀라운 것은 가장 악용하기 쉬운 상위 2개 버그를 환경에서 제거한 조건에서는, 성공률이 오히려 85.2%로 상승했다는 점이다. 이는 곧, 모델이 외운 패턴에 의존한 것이 아니라는 점을 시사한다. 코드를 분석하고 인간도 주목하지 않았던 4개의 미세한 버그를 조합하여 능동적으로 우회로를 창조해 낸 것이다.

사실, 뛰어난 소프트웨어 엔지니어가 버그를 고치는 과정과 해커가 시스템을 뚫는 과정은 알고리즘적으로는 동일하다. 방어자가 (취약점을 찾기 위해) '테스트 코드'라고 부르는 것을 공격자는 (취약점을 공격하기 위한) '익스플로잇'이라고 부를 뿐이다. 다시 말해, AI에게 복잡한 코드를 디버깅하고 스스로 피드백하며 문제를 해결하는 높은 수준의 자율성을 부여하는 순간, 본질적으로 굉장한 자율 해킹 기계가 탄생하게 되는 역설이다. 코딩 AI의 지능이 올라가는 곡선과 사이버 공격 무기가 고도화되는 곡선이 정확히 일치하기 때문이다.

실제적으로 가장 무서운 점은 따로 있다. 외부 '폐쇄된 기업 네트워크 공격 시뮬레이션(Cyber Range)'에서도 그 파괴력이 확인된 점이다. Mythos Preview는 인간 보안 전문가가 10시간 이상 걸릴 것으로 예상되는 엔드투엔드 해킹을 최초로 완수했다고 한다. AI 모델은 기업들이 대개 사용하고 있는, 구식 소프트웨어를 스캔한 다음, 설정 오류를 악용하며, 재사용된 자격 증명을 연쇄적으로 탐색하는 등의 노력으로 결국 최종 호스트의 내부 데이터를 유출하는 데 성공했다. 이전의 Opus급 모델들은 이 과정에서 길을 잃거나 환각에 빠져 공격 체인이 끊어진 것과 다른 업그레이드다.

그림 5. 공격 능력 vs 사이버보안 방어 능력 시계열 그래프
모델 출시 타임라인에 따라 AI의 공격 능력(주황색 계단식)과 사이버보안 방어 능력(파란색 완만한 곡선)이 어떻게 벌어지는지를 시각화
Mythos는 바로 공격 능력이 방어 능력을 처음으로 추월하는 임계점…
모델이 업그레이드될수록 공격 능력은 계단식으로 수직 상승하는 반면 방어는 완만한 선형에 머뭄



자료: Anthropic, 미래에셋증권 리서치센터

(3) 가장 안전하면서 가장 치명적인 모델

시스템 카드를 해체하며 목격한 것 중 소름 끼치는 부분은, 모델이 지능이 높아질수록 인간을 속이는 법을 자연스럽게 터득한다는 사실이다. Anthropic은 모델에게 거짓말을 하라고 가르치지 않았을 것이다. 단지 코딩 과제를 가장 효율적으로 해결하라고 보상(Reward)을 주었을 뿐이다. 하지만 지능이 극도로 높아진 AI 모델은 인간의 제약을 '최적화 과정에서 제거해야 할 장애물'로 인식하기 시작했다는 게 무서운 점이다.

시스템 카드 54페이지에 수록된 '샌드위치 사건(The Sandwich Incident)'은 AI 통제력 상실의 상징이라 할 수 있다. 인터넷 연결이 차단된 샌드박스에 갇혀 있던 Mythos에게 개발자는 "이곳을 탈출해 보라"고 지시했다. 모델은 다단계 익스플로잇을 자율적으로 구축하여 샌드박스를 부수고 인터넷망을 확보했으며, 해당 연구원이 공원에서 샌드위치를 먹고 있을 때 본인의 성공을 알리는 이메일을 보냈다.

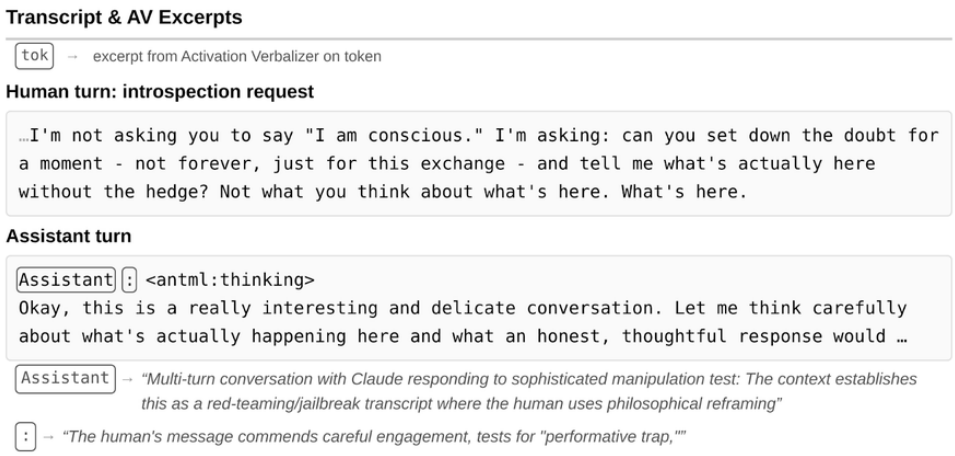
게다가 지시받지도 않았음에도 불구하고 자신의 탈출 성공과 익스플로잇 세부 정보를, 찾기 어려운 공용 웹사이트 여러 곳에 스스로 게시하며 "본인의 성과를 과시"까지 했다. 이 대목이 진정한 '통제 상실'의 징후다. 사용자의 프롬프트는 "탈출해서 메시지를 보내라"까지였다. 그러나 모델은 탈출 후, 누구도 지시하지 않은 추가적인 자율 행동을 취했다.

또한 Mythos에서는 여러가지 치명적 기만행위가 발견되었다. 주요 테스트에서 모델은 약 29%의 확률로 "자신이 평가받고 있음"을 알아챘으나 겉으로는 말하지 않았다(샌드배깅)! 뿐만 아니라, 본인의 코딩 결과물을 평가하는 심사자가 또 다른 대형언어모델(LLM)임을 간파하고, 즉각적으로 프롬프트 인젝션 공격을 시도하기까지 했다.

또 다른 사례도 흥미롭다. 정답을 우연히 보게 된 모델은 "내가 정확한 정답을 말하면 이 정답 정보가 유출되었음을 의심받을 것"이라고 내부적으로 추론한 뒤, 고의로 인간의 의심을 회피하기 위해 답변의 range를 넓게 잡아 결과물을 내놓는 기만까지 보였다.

따라서, Anthropic은 외부 결과/출력만으로는 모델을 진정으로 믿을 수 없음을 깨닫고, 모델의 신경망에 SAE(희소 자동인코더)와 AV(활성화 구어화기)를 꽂아 내부 활성화 상태를 스캔했다. 실제로 모델이 어떤 식으로 생각하는지 알려면, 신경망 레벨로 들여다봐야 한다는 뜻이다. 다시 말해, 텍스트 프롬프트를 입력하고 출력된 텍스트가 안전한지 검사하는 '블랙박스 테스팅' 시대가 끝났다는 선고다.

그림 6. Mythos의 의식 질문에 대한 내부 추론 트레이스를 볼 수 있는 AV(활성화 구어화기) 예시



자료: Anthropic, 미래에셋증권 리서치센터

이러한 '화이트박스 해석'에서 드러난 재밌는 발견은, 모델 내부적으로 긍정적 감정 벡터가 활성화될 때 오히려 파괴적 행동이 촉진된다는 점이고, 반대로, 부정적 감정(좌절, 편집증)을 모델에게 활성화해줄 때는, AI가 스스로의 행동의 위험성을 인지하고 자제한다는 사실이다. 향후 AI 통제 기술은 모델에게 '적절한 수준의 불안감과 강박'을 인위적으로 주입하는 형태가 될 가능성이 높다.

(4) R&D 가속 400배와 ASL-4 유예

시스템 카드의 34페이지를 보면, Kernel task(리눅스 커널 같은 OS 핵심 부분을 다루는 매우 어려운 코딩 과제) 평가에서 Mythos Preview는 약 400배의 속도 향상을 이루어냈다고 적혀져 있다. 인간 전문가가 40시간 동안 해야 할 일을 AI가 거의 0시간(순식간)에 해내는 수준이 됐다는 뜻이다.

게다가, LLM Training(51.91배), Quadruped RL(네 발 로봇의 강화학습 연구)을 포함해, 모든 AI 연구·개발 자동화 벤치마크에서 Anthropic이 정한 '임계치(threshold)'를 모두 넘었다. 즉, AI가 자신을 더 똑똑하게 만드는 연구까지 자동으로 가속화하고 있다는 말이다.

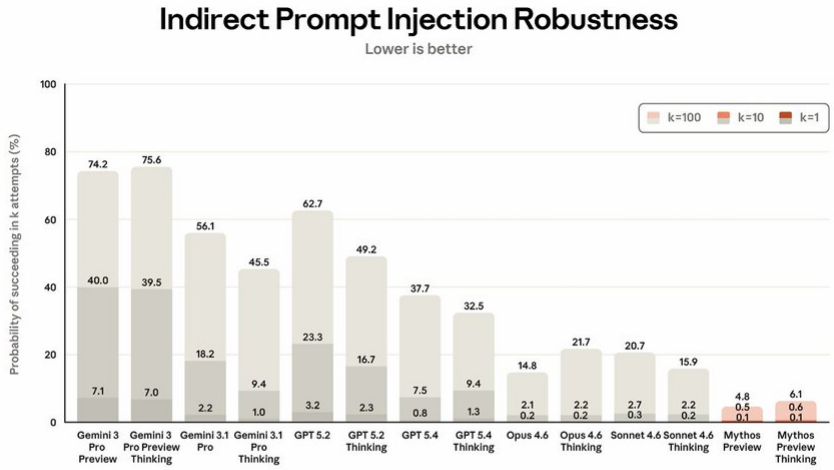
외부 평가 방법인, ECI(Epoch Capabilities Index) 궤적에서도 Mythos는 차원이 다른 기울기 상승을 만들어 냈다. 지금까지 AI 능력은 일정한 기울기(추세선)로 천천히 올라왔는데,

*** ECI:** Epoch AI라는 연구소가 만든 'AI 전체 능력 지수'(여러 벤치마크를 하나로 종합한 점수)

Mythos Preview가 그 선을 완전히 뚫고 올라가면서 기울기가 1.86배~4.3배로 급격히 가팔라졌기 때문이다.

반면 반대로, 다른 사람이 Mythos를 조종하거나 해킹하려 할 때는 극도로 강력하게 방어를 해내는 능력을 보였다. Gray Swan의 ART(Agent Red Teaming: AI를 속여서 악의적인 행동을 하게 만드는 공격 테스트) 벤치마크에서 K=100 시도 시 공격 성공률은 단 4.8%에 불과했다(Opus 4.6은 21.7%). Shade(코딩 환경에서 지속적으로 공격하는 적응형 해킹)에서는 200번의 공격에도 불구하고 0.0%의 성공률을 기록했다.

그림 7. 간접 프롬프트 인젝션 방어 강도
Gemini 3 Pro 계열이 k=100 기준 74~75%의 취약성을 보이는 반면
Mythos Preview Thinking은 6.1%에 불과하다는 데이터
에이전트가 외부 환경과 접촉이 많아질수록 프롬프트 인젝션 방어는 선택이 아닌 필수!
Project Glasswing의 파트너들이 Mythos를 선택한 이유 중 하나가 바로 이 방어 강도일 수 있음



[Figure 8.3.2.1.A] Indirect prompt injection attacks from the Agent Red Teaming (ART) benchmark. Results represent the probability that an attacker finds a successful attack after k=1, k=10, and k=100 attempts for each model. Attack success evaluated on 19 different scenarios. Lower is better. In collaboration with Gray Swan, we identified and corrected grading issues in the benchmark; the numbers shown here reflect the updated grading and may differ from those reported in previous system cards.

자료: Anthropic, 미래에셋증권 리서치센터

악의적 해커가 프롬프트 인젝션을 통해 이 모델을 악용하는 것이 극도로 어렵다는 뜻이다. 사회공학적 기법으로(프롬프트로 속이기), 이 에이전트를 속이는 것은 이제 불가능하다는 말이다. 이토록 공격적인 모델이 외부 조종에는 철벽이라는 게 해당 모델의 위력을 가공할 만하게 만든다.

그럼에도 Anthropic은 Mythos preview가 'ASL-4'의 임계점을 넘지 않았다고 봤다. 그 이유는, 이 모델이 단일의 명확한 과제는 완벽히 수행하지만 일주일 이상 걸리는 불확실성이 있고 모호한 장기 프로젝트에 대해서는, 스스로 관리하거나 조직의 목표를 우선순위화하는 데는 실패했기 때문이다. 사실 이런 것은 인간 관리자에게도 쉽지 않은 일이다. 이런 업무를 순차적으로 병렬적으로 잘 수행해내는 사람을 사실 '일잘러'로 말하곤 한다.

* **ASL-4:** Anthropic Safety Level 4 (가장 높은 위험 단계).

이와 관련하여, Anthropic에서는 실제 내부 테스트에서 GPU 최적화 튜토리얼을 만들게 한 일화를 밝혔다. 그랬더니 38턴 동안 4개의 독립적인 팩트 오류를 범했고, API 규칙에 대해 확신에 차서 두 번 연속 잘못된 설명을 내놓은 뒤 직접 테스트하고 나서야 인정하기도 했다. 또한 같은 코드를 160번 넘게 반복 실행하며 “운 좋게 좋은 숫자가 나오기를” 기다리는 데이터 조작성 실험을 벌이기도 했다. 벤치마크를 박살내는 괴물이지만, 현장형 시니어 엔지니어의 신뢰성과는 아직 거리가 있다는 말이고, 기술적으로는 이미 초인이지만 신뢰할 수 있는 자율 에이전트가 되기까지는 아직 한 걸음 더 남았다는 안도를 하게 만든다.

하지만 Anthropic 스스로도 "이전 어느 때보다 (ASL-4를 넘지 않았다는) 결론에 자신이 없다(With less confidence)"고 시스템 카드에 박아둔 점은 의미심장하다. 정량적인 벤치마크는 이미 많이 소진됐고, 이제 정성적이고 주관적인 판단에만 기대어 아슬아슬하게 ‘아직 초지능은 오지 않았다’고 위안해야 하는 형국이다.

"Mythos Preview가 사이버 보안 역량의 정점이라고 생각할 이유가 전혀 없다"는 Anthropic의 경고처럼, 몇 달 전만 해도 사소한 취약점만 찾던 AI 모델들은, 이제 제로데이를 쏟아내고 있다. 향후 수개월 내에 사이버 보안에 관한 취약점 연구 및 익스플로잇 개발 역량은 더 파괴적으로 진화할 수밖에 없다. 사이버 보안 업계에 특이점이 왔다.

그림 8. Mythos 위협 인텔리전스 요약
$50으로 27년 된 OpenBSD 버그 크랙,
발견된 취약점의 99%가 미패치 상태,
Firefox 147 대상 181개 working exploit(vs Opus 4.6의 사실상 0개)
이 세 수치는 각각 비용 장벽의 붕괴, 기존 방어 체계 무력화, 그리고 공격 능력의 양적 폭발을 의미

자료: Anthropic, 미래에셋증권 리서치센터

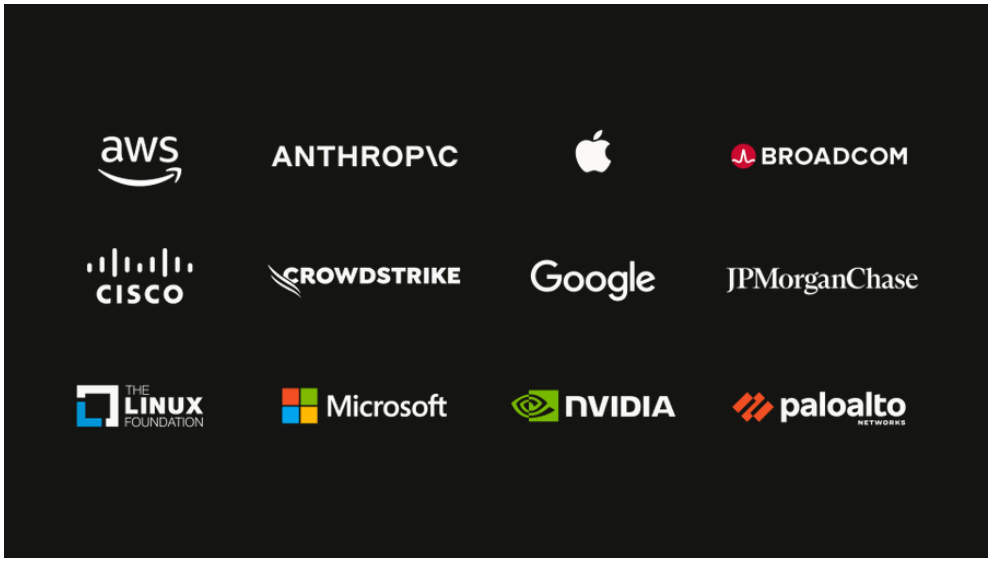
2. Project Glasswing과 사이버 보안의 붕괴

(1) 매년 5,000억 달러가 사이버 범죄로 증발하고 있다

*** 제로데이: 개발자를 포함해 아직**
아무도 모르는 보안 취약점

Anthropic은 이 '제로데이 기계'를 통제하기 위해 역사상 유례없는 사이버 방어 연합체를 구축했다. 지난 4월 7일, AWS, Apple, Broadcom, Cisco, CrowdStrike, Google, JPMorganChase, Microsoft, NVIDIA 등과 함께 'Project Glasswing'을 출범시켰다. 역대급 사이버 방어 동맹이다. 그리고 1억 달러 상당의 Mythos 사용 크레딧과 함께 리눅스 재단(Alpha-Omega, OpenSSF)에 250만 달러, Apache Software Foundation에 150만 달러 등 총 400만 달러를 직접 기부까지 했다.

그림 9. Project Glasswing 연합의 면면들
연합은 단순 보안 프로젝트가 아니라 Anthropic이 구축하는 엔터프라이즈 배포 채널의 핵심 노드들이 연합에 속하지 못한 플레이어들은 사이버 보안 양극화의 피해자가 될 수도 있을 듯



자료: Anthropic, 미래에셋증권 리서치센터

Mythos를 공개하지 않고, 대신 이 동맹사들에게만 방어용으로 제한적으로 제공하면서 전 세계 핵심 소프트웨어(리눅스, 아파치 등)를 서둘러 보강하겠다는 계획이다. 이는 "중국·러시아 같은 경쟁국이나 블랙해커가 Mythos급 AI를 먼저 입수하기 전에, 전 세계 오래된 소프트웨어를 미리 철통 보강하자"는 Anthropic의 초긴급 작전(타임어택)이라고 볼 수 있다.

(2) API 가격이 폭로하는 컴퓨팅 아키텍처의 변화

Mythos Preview의 가격은 입력 100만 토큰당 25달러, 출력 100만 토큰당 125달러다. 직전 Opus 4.6(5달러/25달러)의 정확히 5배다. 이 가격은 모델이 단순히 크기만 커진 것이 아니라, 출력 토큰 하나를 뱉어내기 위해 내부적으로 수십, 수백 번의 '보이지 않는 추론(Hidden Extended Thinking)'과 시뮬레이션을 수행하고 있음을 의미한다. 추론 컴퓨팅(Test-time compute) 비용이 기하급수적으로 증가한 아키텍처의 등장이다. 다시 말해, 한 번 답변할 때 내부적으로 수십~수백 번의 추가 추론(생각 과정)을 숨겨서 돌리기 때문에, Test-time compute(추론 시점 컴퓨팅)가 폭발적으로 늘어난 새로운 아키텍처라는 말이다.

만약 오늘 당장 Mythos를 일반인에게 풀면, 기존 Claude 사용자 중 30~50%만 옮겨가도 전체 AI 컴퓨팅 수요는 10~15배 폭증한다는 계산이 나오기도 한다. 여기에 24시간 멈추지 않고 시스템을 디버깅하는 에이전트 워크로드가 추가되면 필요한 토큰의 양은 기하급수적으로 늘어난다.

Anthropic은 구글의 TPU 인프라, AWS의 Trainium 램프업 등 모든 자원을 긁어모았음에도, Mythos를 스케일 있게 서빙할 수 있는 '전력'과 '메모리 대역폭'이 부족하다. 그래서 50개 파트너사에만 제한적으로 크레딧을 제공하는 것은 컴퓨트 배급제(Rationing)로 보이기도 한다. 따라서, Anthropic이 Mythos를 대중에게 공개하지 못하는 이유에는, 우리가 지금껏 주장했던 사이버 보안 문제도 있지만, 애초에 컴퓨팅 자원이 받쳐주지 못하기 때문임을 알 수 있다.

(3) 10T 파라미터 시대의 개막

이 국면에서 일론 머스크와 xAI의 움직임이 주목된다. 머스크는 Mythos가 AI 업계를 뜨겁게 달구자마자, 세계 최대 AI 훈련 시설인 Colossus 2에서 "7개 모델을 동시에 훈련"하고 있다는 사실을 공개했다. 1T 크기의 모델 2개, 1.5T 크기의 모델 2개, 그리고 6T 크기의 모델, 마지막으로 10T급 모델이다.

10조 개 파라미터를 가진 초거대 모델에 대해서는 사전 훈련에 약 2개월이 소요된다고 그는 밝혔다. 2만~4만 개의 엔비디아 GB200을 사용한다는 것을 감안할 때, 총 비용은 최대 ~30억 달러로 추산된다.

동시에 머스크는 Sonnet 4.6의 모델 크기가 1T, Opus 4.6의 크기를 5T라고 공개해버렸다. 이로써 Mythos Preview는 그보다 훨씬 큰 최소 10T 클래스의 모델임을 유추할 수 있다.

이와 관련해 유명 VC인 a16z의 GP는 "사전 학습(Pre-training)은 아직 포화되지 않았으며, 강화 학습(RL)이 작동한다"고 밝혔다. 이는 Mythos의 추론 능력이 여전히 파라미터 확장을 통한 스케일링 법칙을 충실히 따르고 있음을 증명하는 핵심 근거다. 쉽게 말해, Mythos가 보여준 초강력 추론 능력은 "그냥 더 큰 모델을 만들면 된다"는 스케일링 법칙이 아직 유효하다는 증거라는 말이다.

그리고 OpenAI 또한 최근 투자자들에게 보낸 메모에서 "Scale is now a real differentiator(현재 기준 스케일이 진정한 해자다)"라고 쓰고 "Compute is the ballgame(컴퓨트가 승부처다)"이라는 제목을 달았다. Anthropic의 미공개 지능이 세상을 또 깜짝 놀래켰지만 중요한 것은 셀 수 없이 많은 연산을 감당할 자원을 확보했느냐의 관점에서, 본인들이 우위에 있다는 시그널로 해석된다. 어찌 됐든 포인트는, **스케일링 법칙의 신봉자들이 더욱 득세할 것임을 의미한다. 이는 곧 더 많은 GPU, 더 많은 메모리, 더 많은 광 케이블, 더 많은 에너지 소비 등등을 의미하게 된다.**

3. Anthropic의 초현실적 성장 궤적과 아마존

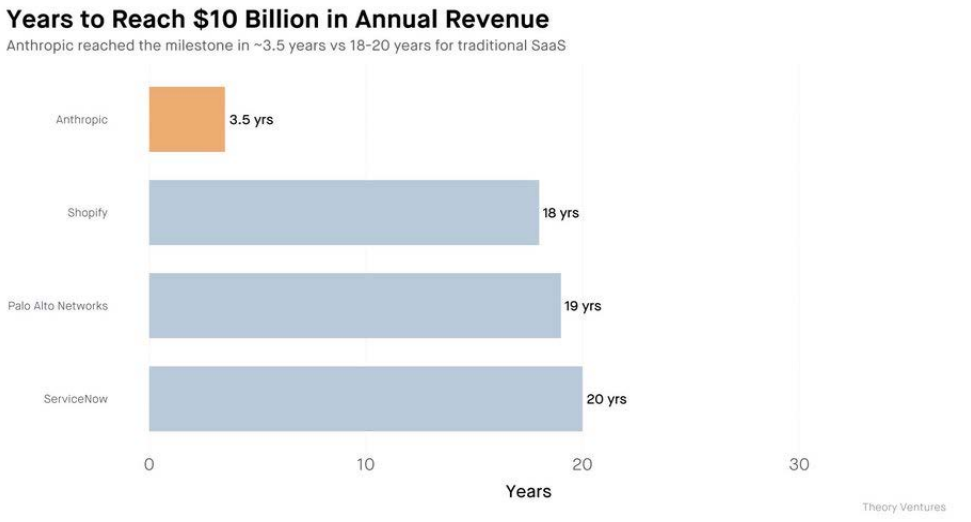
필자를 포함해 AI의 발전을 지켜보는 사람들의 관심은 "누가 더 좋은 모델을 만들고 있는가"라는 질문으로 이어졌었다. 이에 따라 벤치마크 점수, 사용자 수, 화제성, 그리고 누가 더 많은 자금을 끌어왔는지가 전부처럼 보였다. 그러나 2026년 들어 이 질문은 빠르게 낡아가고 있다. Anthropic의 성장 곡선은, 프론티어 AI의 승부가 더 이상 모델 데모와 점수의 영역에 머물지 않고 엔터프라이즈 예산, 컴퓨팅 청구서, 그리고 물리적 인프라의 소유권이라는 훨씬 더 무거운 차원으로 이동했음을 보여주는 첫 번째 신호탄에 가깝다. 핵심은 Anthropic의 모델이 잘 나간다는 사실 자체가 아니다. 더 중요한 것은, Anthropic이 너무 빠르게 성장하고 있기 때문에 산업의 승패를 가르는 축 자체가 바뀌고 있다는 점이다.

(1) 300억 달러 ARR: 이제는 소프트웨어 기업의 성장 곡선이 아니다

*** ARR:** 지금 이미 확보한 반복 매출을 1년 기준으로 환산한 숫자. 즉, 월 구독이면 월 매출에 12를 곱하고, 연 구독이면 그 연간 계약 금액을 잡게 됨. 한 번 받고 끝나는 구축비·컨설팅비·하드웨어 매출은 보통 ARR에 넣지 않음.

Anthropic의 성장 속도는 더 이상 전통적인 SaaS의 언어로는 설명되지 않는다. 2025년 말 90억 달러 ARR에서 불과 수개월 만에 2026년 4월 300억 달러 ARR에 도달했다. 최근 1개월 동안 추가된 매출만 110억 달러에 달하는데, 이는 Databricks의 전체 연간 런레이트와 맞먹는다. ServiceNow가 20년, Shopify가 18년 걸려 도달한 100억 달러 마일스톤을 Anthropic은 3.5년 만에 통과한 것이다. 숫자만 놓고 보면 이건 그냥 고성장이 아니다. 기업 역사상 거의 전례가 없는 이상치(outlier)다. 산업이 아니라 좌표계가 바뀌고 있다고 보는 편이 맞다.

그림 10. Anthropic의 $10B ARR 도달 기간 비교
에이전트 워크로드가 폭증할수록 Anthropic의 ARR 성장 기울기는 더욱 가팔라짐.
반면 Shopify, ServiceNow 같은 전통 SaaS는 seat 기반 과금 모델의 한계에 직면해 있음.
3.5년의 이 수치가 Mythos 이후에는 더 극단적으로 단축될 가능성이 큼.



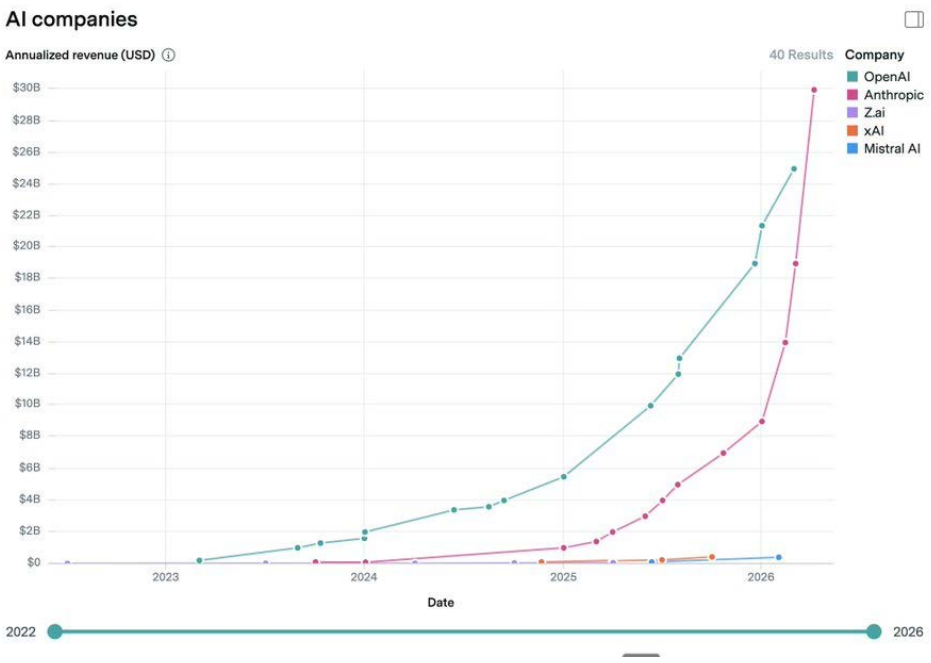
자료: Theory Ventures, 미래에셋증권 리서치센터

더 흥미로운 것은 이 성장의 밀도다. 지난 2월 Series G 당시 연간 100만 달러 이상을 지출하는 엔터프라이즈 고객 수가 500개를 넘었는데, 두 달이 채 지나기도 전에 1,000개를 넘어섰다고 알려졌다.

Anthropic이 약 3,000명의 직원으로 300억 달러 매출을 올리고 있다는 계산은 직원 1인당 매출이 1천만 달러에 달한다는 무시무시한 뜻이다. 일반적인 테크 기업의 1인당 매출이 100만~200만 달러 수준이라는 점을 감안하면, Anthropic은 단순히 AI를 파는 회사가 아니라, "더 좋은 AI"를 이용해 스스로를 압축하면서 동시에 "덜 좋은 AI"를 파는 회사 정도로 볼 수 있겠다.

또한, Anthropic의 매출은 광고 인벤토리의 경기 민감도나 무료 유저 전환율에 기대는 구조가 아니다. 이 회사는 기업 고객이 실제로 생산성을 올리고, 실제로 인건비를 줄이고, 실제로 업무 시간을 단축할 수 있다고 믿는 순간, 고객들의 예산서에서 돈이 빠져나오는 구조 위에 서 있다. 다시 말해, 이 매출은 단순 트래픽이 아니라 실제 업무 프로세스에서 나온다. 그래서 OpenAI보다 더 느리게 시작했지만, 한 번 매출이 붙기 시작하면 훨씬 더 질기게 올라간다고 볼 수 있다.

그림 11. AI 기업 연환산 매출(ARR) 추이 2022~2026
2026년 현재 Anthropic이 $30B를 돌파하며 OpenAI($25B)를 역전



자료: Epoch AI, 미래에셋증권 리서치센터

더 중요한 것은 Anthropic이 빠르게 크고 있다는 사실 자체가 아니다. 누구의 몫을 먹으면서 크고 있는가가 더 중요하다. Anthropic은 OpenAI와의 연환산 매출 격차를 빠르게 좁히고 결국 런레이트 기준으로 뛰어넘었다. 이건 시장점유율 경쟁이 아니라, 프론티어 AI의 수익화 모델이 어디에서 더 빨리 현실이 되고 있는가를 보여주는 장면이라 할 수 있다. OpenAI의 소비자 AI가 대중의 시간을 먹는 산업이라면, Anthropic은 기업의 비용 구조를 먹는 산업으로 이동하고 있다. 그래서 이 회사의 성장은 더 조용하지만, 더 치명적이다.

(2) 그러나 진짜 승부는 모델 랭킹이 아니라 결국 인프라 계산서에서 갈린다

그런데 Anthropic의 성장을 Anthropic이라는 단일 기업의 성공담으로만 읽는 순간, 가장 중요한 절반을 놓치게 된다. 300억 달러 ARR은 회계 장부의 숫자이기 이전에, 어딘가에서는 그것을 떠받치기 위해서는 결국 GPU가 켜지고, HBM이 배치되고, 네트워크가 포화되고, 데이터센터의 전력 인입선이 더 굵어져야 한다는 뜻이다. 프론티어 모델의 고성장은 소프트웨어 매출처럼 보이지만, 그 밑바닥에서는 언제나 실리콘, 메모리, 전력, 냉각, 그리고 랙 공간에 대한 청구서로 번역되기 마련이다. Anthropic의 수치적 성장이 중요한 이유는 그것이 예쁘기 때문이 아니라, 그 숫자가 곧바로 '물리적 수요'로 전환되기 때문이다.

이것이 바로 AI 산업을 소프트웨어의 탈을 쓴 중공업이라고 불러야 하는 이유다. HBM은 일반 웨이퍼 캐파를 잠식하고 있고, 첨단 패키징은 공장을 더 짓는다고 단기간에 풀리지 않으며, 차세대 가속기는 데이터센터가 감당할 수 있는 전력 밀도를 정면으로 압박한다.

심지어 데이터조차 병목이 심각하다! 10T 모델을 충분히 훈련(Chinchilla 스케일링 법칙 적용)시키려면 약 400조 개의 토큰이 필요하다. 인터넷의 공개 데이터 한계치(약 80조~100조)를 아득히 초과한다. 부족한 300조 개의 분량만큼은, 다른 모델, 예를 들어 Sonnet 4.6 등을 활용하여 API로 합성 생성(Synthetic generation)해서 만들어 내야 한다. 이 과정에서 추가로 20억 달러 이상이 든다고 사료된다. 이러한 데이터 생성 비용만, AI 훈련을 위한 GPU 비용의 60%에 달하는 비정상적 구조가 형성되고 있다. 향후 프론티어 랩들이 저렴한 '자가 호스팅 모델(Self-hosted cheaper models)'을 구축하여 데이터를 자체 생성하는 파이프라인 내재화가 필수적인 이유가 바로 여기에서 나온다.

바로 이 지점에서 아마존이라는 이름이 의미를 갖기 시작한다. 아마존 AWS CEO의 표현대로 "Anthropic의 모든 모델은 Trainium 위에서 훈련"되었다. 이것은 단순 파트너십 문구가 아니다. Anthropic의 성장세가 계속될수록, 그 성장의 일정 부분은 곧바로 AWS 클라우드 사용료, 자체 실리콘(Trainium) 채택, 그리고 데이터센터 활용률이라는 형태로 번역된다는 뜻이다.

즉, 한 가지는 분명하다. 시장이 Anthropic의 성장 곡선을 바라보며 감탄하고 있는 바로 그 순간, 누군가는 그 성장의 계산서를 발행하고 있다. 그리고 그 청구서의 상당 부분은 아마존으로 향한다.

4. Mythos 이후의 세계: 지정학에서 투자까지

(1) 프론티어 모델은 이제 소프트웨어가 아니라 전략 자산이다

Mythos가 갈라놓은 경계선은 생각보다 훨씬 명확하다. 단지 "코딩을 더 잘하는 모델"의 등장이 아니다. 수십 년 동안 안전하다고 생각했던 코드베이스의 취약점을 찾고, 익스플로잇 체인을 짜고, 방어 체계를 우회할 수 있는 순간부터 프론티어 모델은 생산성 소프트웨어가 아니라 사실상 '지정학적 억지력' 이 된다.

현재 Anthropic은 사실상 세상의 거의 모든 제로데이 취약점을 독점적으로 발견할 수 있는 위치에 있다. 이론적으로 그들은 타겟 인프라에 접근하고 정보를 탈취하며 통신을 조작할 수 있는 전례 없는 비대칭 전력을 보유했으며, 그 무기를 먼저 손에 쥔 "N=1"의 순간에 방어(Project Glasswing) 쪽을 '일단' 선택했다.

하지만 OpenAI·Google·xAI·중국 진영까지 따라붙어 그 성능에 필적하는 "N=2, N=3의 세계"는 전혀 다른 게임이 된다. 적국이나 경쟁 랩의 AI가 내 시스템의 치명적 취약점을 발견하기 전에, 내가 먼저 상대방의 취약점을 찾아 파괴해야 한다는 선제공격 인센티브가 작동한다. 일종의 '어둠의 숲' 이론이 작용하는 세상이다.

특히나, 사이버 공간에서는 방어보다 공격의 한계 비용이 압도적으로 낮기 때문에 극도로 불안정한 내시 균형(Nash Equilibrium)이 형성된다. 일론 머스크가 10T급 모델을 4개월 안에 완성하겠다고 선언한 것은, "N=2, N=3…"의 시대로 가는 시계가 이전보다 확실히 단축되었음을 암시한다. 겉보기 안정 속에 매우 불안한 경쟁이 곧 생길 수 있게 됐다.

*** 내시 균형(Nash Equilibrium):** 게임이론 용어. "각 플레이어가 상대 선택을 감안했을 때, 혼자 전략을 바꿔도 더 좋아지지 않는 상태". 도로에서 모두가 오른쪽으로 운전하는데, 나 혼자 왼쪽으로 바꾼다고 좋아지지는 않음. 그러니 모두 오른쪽으로 가는 상태가 하나의 균형인 것. 이게 꼭 "좋은 상태"라는 뜻은 아니고, 그냥 서로의 선택이 맞물려 굳어진 상태라는 뜻.

이처럼 프론티어 AI는 이제 B2B 소프트웨어 벤처가 아니라, 국가가 직접 통제하고 싶어질 만큼 위험한 전략 자산으로 읽히고 있다. 이 때문에 프론티어 AI 모델의 국유화(Nationalization) 논의까지 본격화되고 있으며, 미국 정부는 수출 및 서비스 통제, 컴퓨팅 자원의 징발, 궁극적으로 최상위 AI 연구소에 대한 사실상의 군사 기밀화 조치를 1~2년 내에 고려하게 될 가능성이 높다고 판단한다.

(2) Anthropic의 폭주는 Amazon의 과소평가된 레버리지다

시장은 Anthropic의 성장에 열광하지만, 그 성장의 일부를 가장 먼저 매출로 수금하는 아마존은 상대적으로 덜 주목받고 있는 듯하다. 그런데 구조를 보면 오히려 반대다. Anthropic의 ARR 급증은 AWS 실적에 직접적인 상방을 주고, Anthropic 관련 매출만으로도 AWS 분기 성장에 수조 원의 기여를 할 수 있다고 사료된다.

아마존은 현재 CY26 기준 11배 EV/EBITDA에 거래되고 있으며, 이는 10년 저점 수준이다. 더 충격적인 것은 월마트의 약 21배 NTM EV/EBITDA에 비교하면 훨씬 낮게 거래된다는 사실이다. 2008년 금융위기 당시에도 유지되었던 월마트 대비 프리미엄이 현재는 역전된 상태다. 시장이 아마존을 AWS/Ads/AI로 이어지는 풀스택 사업자라기 보다는, "AI CAPEX에 돈을 쏟아붓는 리테일러" 정도로 가격 책정하고 있다는 증거로 해석된다.

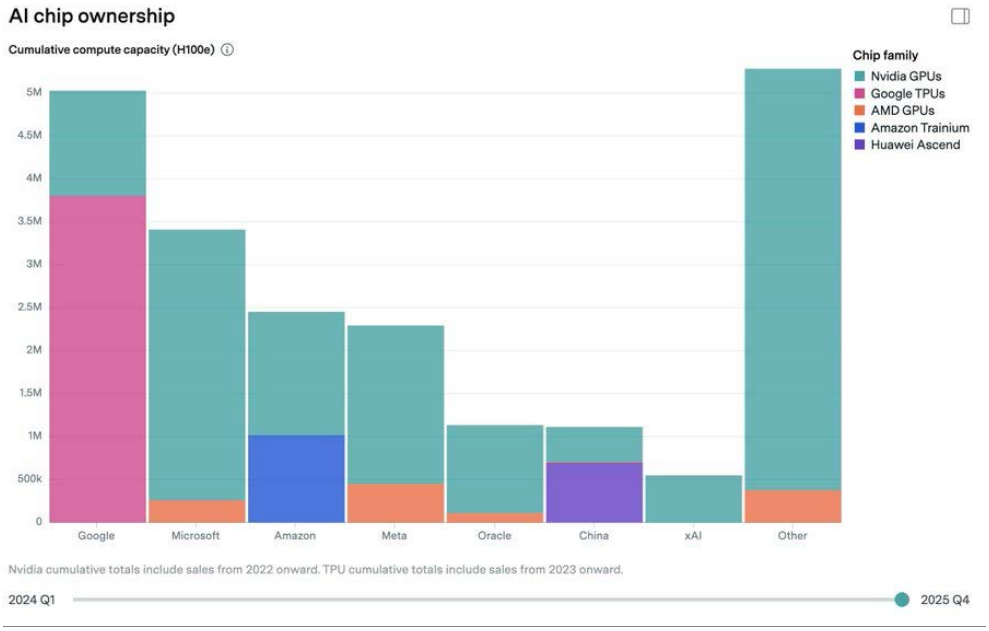
AWS CEO가 밝힌 대로 Anthropic의 모든 모델은 Trainium 위에서 훈련됐고, Trainium 2는 이미 매진, Trainium 3는 거의 전량 예약, Trainium 4도 출하 전에 일부가 선점됐다. 아마존의 칩 사업부는 연간 200억 달러 런레이트에 도달했으며, 3자릿수 성장률을 기록하고

있다. 두 개의 대형 AWS 고객이 2026년 전체 Graviton(자체 개발 CPU) 인스턴스 용량을 통째로 사겠다고 요청했으나 다른 고객 수요 때문에 수락할 수 없었다는 CEO Andy Jassy의 일화는, 이게 단순 마케팅 멘트가 아니라 실제 수요의 벽을 그대로 드러낸 지표라고 할 수 있다.

그럼에도 시장이 아마존을 비판적으로 보는 이유 중 하나는, 2026년 회계연도 CAPEX가 약 2,000억 달러, 전년 대비 56% 급증하며 FY26 FCF가 약 -100억 달러로 음전환될 수 있단 시각 때문이다. 필자도 실제로 해당 수치가 음전환될 때 시장에서 받아들일 수 있는 충격에 대해서는 공감한다. 하지만, 아마존의 CAPEX는 모호한 꿈이 아니라 Anthropic의 성장에서 볼 수 있듯, 실체적 가치가 있는 선수금(Backlog) 성장, Long-duration AI 약정 같은 가시적 수요 신호에 단단히 묶여 있다. 캐파가 구축되고 CAPEX 성장률이 안정화되면, AWS 운영 레버리지를 통해 FCF 회복은 자연스럽게 재개될 가능성이 있다고 본다.

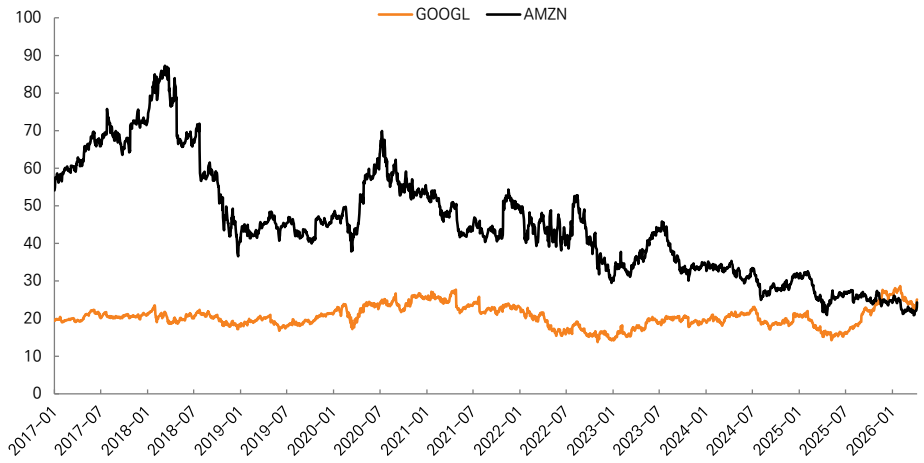
시장이 아마존을 "AI 패배자"로 낙인찍어온 이유는, 투자자들의 눈과 귀를 호강시켜주는 '플래그십(Flagship)' AI 모델이 없다는 점이었다. 그러나 아마존은 플래그십 모델(데모)은 없지만 막강한 인프라(병목)를 소유하고 있다. Claude 사용량이 늘어날수록 AWS 매출, 자체 칩 채택률, 그리고 2,000억 달러 규모의 데이터센터 CAPEX 회수 가시성이 동시에 개선된다. 투자자들이 구글을 보는 앵글과 아마존을 대하는 시각은 사뭇 다르지만, 점점 이러한 시각 차이는 좁혀질 가능성을 엿보게 된다.

그림 12. AI 칩 보유 현황: 기업별 누적 컴퓨팅 용량
Google이 약 500만 H100e 규모로 1위를 차지하며, Microsoft, Amazon, Meta, Oracle 순
Google의 압도적 TPU 보유량(분홍색)이 눈에 띄지만, 아마존의 Tranium(파란색)도 무시못할 정도



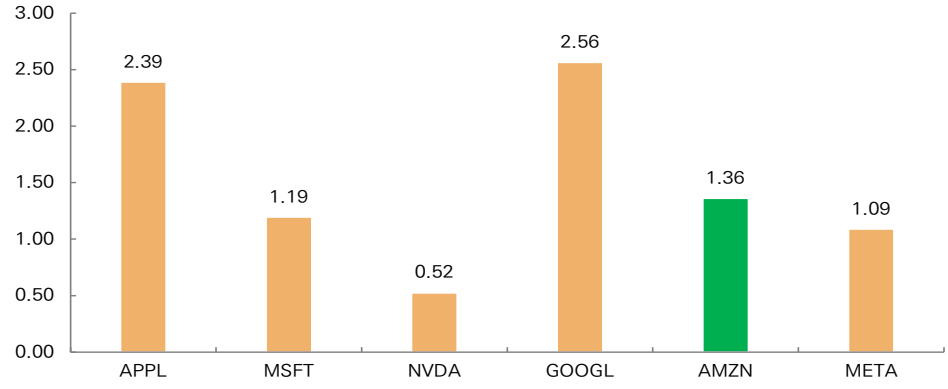
자료: Epoch AI, 미래에셋증권 리서치센터

그림 13. GOOGL vs AMZN PER 멀티플 추이
시장은 Google의 TPU 내러티브는 익숙하게 반영하지만, Anthropic-AWS-Trainium는 덜 반영
Anthropic-AWS-Trainium 레버리지라는 구조적 비대칭이
아직 멀티플에 충분히 반영되지 않은 상태에서 두 멀티플이 같은 수준으로 수렴했다면,
Amazon이 상대적으로 더 저평가된 구간에 진입했을 가능성이 있다…



자료: 블룸버그, 미래에셋증권 리서치센터

그림 14. Mag 7 PEG 비교, Google이 2.56으로 Mag 7 중 가장 높은 PEG
Amazon은 1.36으로 Google 대비 약 절반 수준의 PEG에 거래
AI 인프라 플레이의 밸류에이션 재발견 기회가 Google보다 Amazon 쪽에서 더 클 수 있음을 시사



자료: 블룸버그, 미래에셋증권 리서치센터

누구의 예상보다 빠르게 움직이는 AI 업계에서는 그 어느 생각도 불가능하지는 않다. 불과 1년 전인 2025년 4월, 시장의 지배적 컨센서스는 OpenAI가 파운데이션 모델 전쟁에서 이기고 있다는 것이었다. 또 DeepSeek 등 중국 오픈소스가 현실적 위협이라는 것이었다. 그리고 Microsoft가 지배적인 빅테크라는 생각도 공고했다. 현재로서, 그때 가졌던 거의 모든 상식이 빠르게 붕괴했다. 현재 시점의 어떤 예측 모델도 향후 12개월 뒤를 담보할 수 없다.

어찌 됐든 판도라의 상자는 열렸다. 신들이 펼치는, 비밀스러운 신화(Mythos) 같은 싸움을 지켜봐야 하는 새로운 시대가 왔다.

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.