Market Insight | 2026.2.5

그림: Midjourney(소프트웨어의 껍데기(UI)와, 그 뒤에서 실제로 돌아가는 AI의 괴리)

MIRAE ASSET 미래에셋증권

AI Bi-Weekly

SaaS… 스트레스 많이 받을거야

한종목
chongmok.han@miraeasset.com

김은지
eunji.kim.a@miraeasset.com

CONTENTS

Highlight of the Week 3
SaaS의 종말과 'Deep SaaS'의 부상 3
1인 1자비스 시대? 4

I. AI & Software 5
1. 왜 올해 갑자기 에이전트로의 전환이 크게 체감될까? 5
2. 충격적인 Claude Code 팀 업무 방식 7
3. AI 엔터프라이즈 시장은 구조적 붐을 가리킨다 9
4. Anthropic 쇼크와 SaaS의 위기 13
5. SaaS 종말론에 대한 반론 16
6. 종말은 없다, 'Deep SaaS'의 탄생이 있을 뿐 18

II. AI Agent Issue 24
1. 손발이 달린 에이전트들 24
2. 기반 모델에의 함의 28
3. 하드웨어 시장의 양분화: 더 가볍게, 더 무겁게 30
4. 투자 시사점 32

III. AI Peer Table 33

IV. Appendix: Charts 35

Highlight of the Week

SaaS의 종말과 'Deep SaaS'의 부상

2026년은 AI 역사에서 기술적 특이점을 향한 폭발적 가속의 원년으로 기록될 것입니다. Anthropic이 정의한 대로, 이제 AI는 인간의 개입 없이 스스로 코드를 분석하고 개선하는 '자기개선 루프'를 완성해나가고 있습니다. 이는 기술 발전의 속도가 기하급수적 단계로 전환되는 임계점을 넘었음을 의미하며, 이 통제 불가능한 속도는 기업의 생존 방식을 송두리째 뒤흔들고 있습니다. 이러한 변화 속에서 개발의 패러다임은 완전히 전복되었습니다. 이제 생산성의 병목은 코드 작성이 아니라 쏟아지는 결과물을 검토하는 인간의 '리뷰 속도'에 달려 있습니다.

기업 시장 역시 실험 단계를 넘어 실질적인 변화를 맞이하고 있습니다. 2025년을 기점으로 기업들의 AI 관련 지출은 단순 R&D 예산이 아닌 핵심 운영 비용과 인프라 투자로 전환되었습니다. 엔비디아의 칩과 그 위에서 돌아가는 애플리케이션의 투자 비중이 균형을 이루며 AI가 실제 수익을 창출하는 단계에 진입한 것입니다. 특히 기업들은 자체 모델을 밑바닥부터 만드는 무모한 시도 대신, 검증된 외부 모델을 도입하되 자사의 워크플로우와 데이터를 결합해 내재화하는 'Buy to Build' 전략을 통해 실질 성과를 내는 데 집중하고 있습니다.

하지만 빛이 강하면 그림자도 짙은 법입니다. AI가 코딩 비용을 '0'에 수렴시키면서 소프트웨어 공급이 폭발했고, 이는 지난 15년을 지배해 온 기존 SaaS 비즈니스 모델에 치명적인 위협이 되고 있습니다. 누구나 저렴하게 맞춤형 소프트웨어를 만들 수 있는 시대에, 단순 기능 중심의 서비스나 LLM 래퍼들은 설 자리를 잃고 있습니다. 더욱이 AI 에이전트가 인간의 업무를 대체하면서 '사람 수(Seat)' 기반의 과금 모델이나 화려한 UI로 사용자를 붙잡아두던 방식은, 인간 대신 AI가 백그라운드에서 일하는 시대에 더 이상 유효하지 않게 되었습니다. 시장은 이제 단순한 도구를 파는 기업을 외면하기 시작했습니다.

그러나 이것이 소프트웨어 산업의 종말을 의미하는 것은 아닙니다. 오히려 시장은 '기능'이 아닌 '결과'와 '책임'을 파는 'Deep SaaS'를 중심으로 재편되고 있습니다. 확률적으로 작동하는 AI 에이전트가 제멋대로 행동하지 않도록 기업의 데이터와 자산을 디지털 트윈으로 정의하고 통제하는 팔란티어의 '온톨로지'가 대표적인 예입니다. 또한, 반도체 설계처럼 물리적 법칙과 절대적 정답을 다루는 시놉시스와 같은 기업들은 AI가 대체할 수 없는 '진실의 원천'을 쥐고 있습니다. 이들에게 AI는 위협이 아니라 시스템의 시뮬레이션을 수만 번 반복해 주는 최고의 헤비 유저이며, AI가 고도화될수록 이들의 해자는 더욱 깊어집니다.

결론적으로 우리는 지금 소프트웨어 시장의 거대한 분기를 목격하고 있습니다. 인간의 입력을 돕는 얇은 층의 UI 중심 소프트웨어는 도태되거나 상품화되겠지만, 기업의 원천 데이터를 장악하고 AI 에이전트의 행위를 검증하며 통제하는 플랫폼은 필수 인프라로서 그 가치를 재평가받게 될 것입니다. 결국 화려한 껍데기는 사라지고, 본질적인 시스템과 진실을 다루는 기업만이 살아남아 더 큰 부가가치를 창출하는 새로운 시대가 도래한 것입니다.

1인 1자비스 시대?

2026년 1월, 오스트리아 개발자 Peter Steinberger가 만든 오픈소스 프로젝트 OpenClaw는 GitHub에서 3일 만에 10만 스타를 돌파하며 역사상 가장 빠른 성장을 기록했고, 이를 기반으로 한 AI 에이전트 전용 소셜 네트워크 Moltbook에는 160만 개 이상의 에이전트가 활동 중입니다. 같은 시기 Google DeepMind는 비코딩 DNA 영역의 유전자 조절 기능을 예측하는 AlphaGenome을 Nature에 발표했으며, 텍스트나 이미지로부터 실시간 상호작용이 가능한 가상 세계를 생성하는 Genie 3를 출시했습니다.

세 가지 사례의 공통점은 AI가 조언자에서 실행자로, 도구에서 참여자로 역할을 확대하고 있다는 것입니다. OpenClaw의 에이전트는 사용자 대신 직접 이메일을 보내고 파일을 정리합니다. AlphaGenome은 연구자들이 발견하기 어려운 유전자 발현 패턴을 예측합니다. Genie 3는 게임 개발을 보조하는 대신 세계 자체를 생성합니다. 이러한 전환의 핵심에는 LLM의 도구 사용 능력, 즉 브라우저 제어, 파일 시스템 조작, API 호출, 물리 시뮬레이션 등이 있습니다.

에이전트 시대의 도래는 기반 모델 개발사에게 두 가지 경제적 함의를 갖습니다. 첫째는 API 수익의 구조적 증가입니다. 24시간 상시 작동하는 에이전트는 기존 챗봇 대비 토큰 소비량을 폭증시키며, 현재 160만 에이전트 기준으로도 Anthropic에 연간 1.4~3.4억 달러의 추가 수익을 제공할 것으로 추산됩니다. 둘째는 구형 모델의 내용연수 증가입니다. 에이전트 용도로는 최신 frontier 모델이 아닌 Sonnet이나 Haiku급 모델로도 충분한 경우가 많아, 모델 포트폴리오 전체의 가치가 재평가될 수 있습니다.

하드웨어 시장에서는 AI 컴퓨팅 수요가 데이터센터를 넘어 개인 시장으로 확산되고 있습니다. OpenClaw가 촉발한 Mac Mini 품귀 현상이 이를 상징적으로 보여줍니다. 수요는 저사양과 초고사양으로 양분화되는 특성을 보입니다. API 호출 기반 에이전트 운영을 위한 미니 PC 수요와 로컬 LLM 구동을 위한 고성능 GPU 수요가 동시에 증가하고 있습니다.

투자 관점에서는 도구 사용 능력에 강점을 보이는 Anthropic 등 기반 모델 개발사에 대한 재평가가 필요하며, B2C AI 수요를 반영한 제품 라인을 갖춘 반도체 및 PC 관련 기업들이 수혜를 받을 가능성이 높습니다. AI가 조언자에서 실행자로 전환하는 흐름은 이제 막 시작되었으며, 그 경제적 파급효과는 앞으로 더욱 빠르게 확산될 것입니다.

표 1. AI 관련 주요 일정

일	월	화	수	목	금	토
1	2	3	4	5	6	7
	· PLTR 실적	· AMD 실적	· GOOGL 실적 · ARM 실적	· ILMN 실적 · BE 실적 · AMZN 실적(잠)	· 네이버 실적 · 위스트론 실적(예)	
8	9	10	11	12	13	14
	· 크래프톤 실적	· SMIC 실적 · NET 실적	· SHOP 실적 · VRT 실적	· 소프트뱅크 실적 · HD현대 실적 · COIN 실적	· CRWV 실적(예)	
15	16	17	18	19	20	21
		· CDNS 실적(잠)			· BABA 실적(예)	

자료: Bloomberg, 미래에셋증권 리서치센터

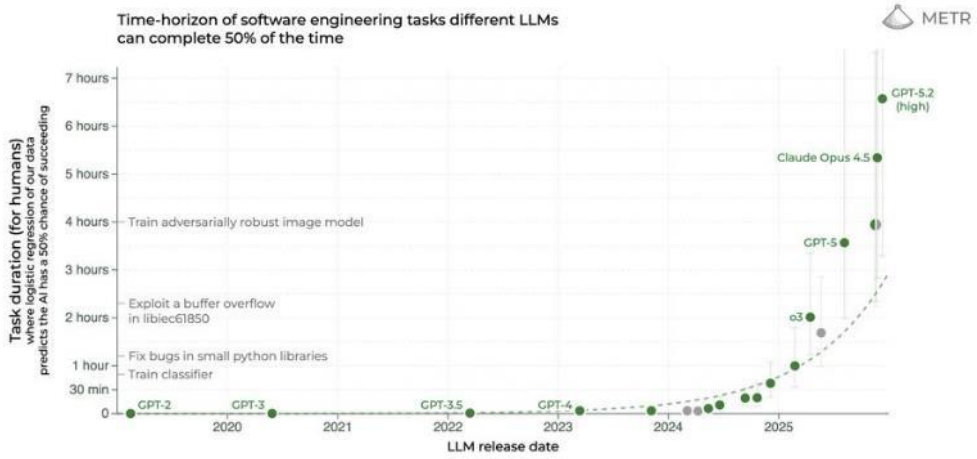
I. AI & Software

1. 왜 올해 갑자기 에이전트로의 전환이 크게 체감될까?

(1) 역사적 임계점(Threshold)의 통과

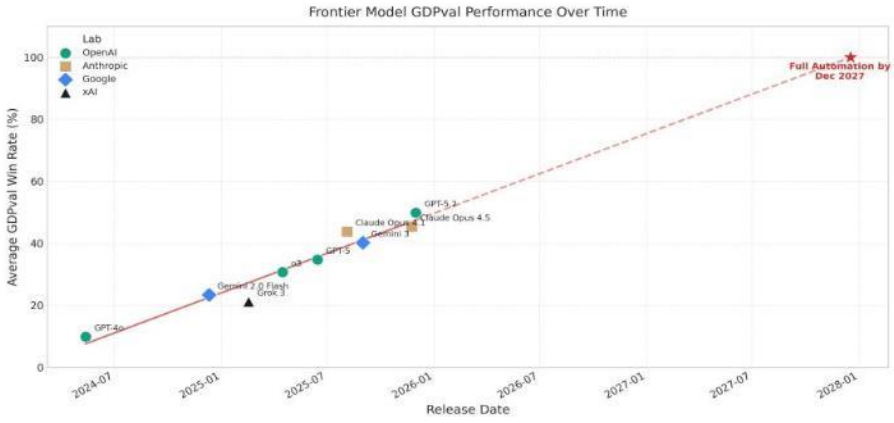
2026년은 인공지능 역사에서 기술적 특이점(Singularity)을 향한 폭발적 가속의 원년으로 기록될 가능성이 높다. Anthropic은 2026년을 단순한 기술 발전의 연장선이 아닌, 인류 역사상 처음으로 자기개선(Self-Improvement)하는 사이버-물리 시스템이 가능해지는 임계점으로 규정했다. AI 발전의 패러다임이 인간에서 AI 주도로 넘어가는 질적전환을 의미한다.

그림 1. METR 50%-time-horizon은 현재 약 6.6 시간
time-horizon은 "인간이라면 연속 '몇 시간'이 걸릴 작업을 AI는 50% 확률로 끝낼 수 있는지 측정"
2024년까지 거의 바닥에 붙다가, 현재 시간지평이 '하키스틱'처럼 급등하는 모습



자료: METR, 미래에셋증권 리서치센터

그림 2. GDPval 추세만 보면, 2027년 12월이 되면 완전 자동화가 될 수도 있다
에이전트의 '업무 수행 능력(=사람 대체)'이 시간축에서 빠르게 임계점으로 접근



자료: Parker Whitfill(MIT 경제학 박사), 미래에셋증권 리서치센터

지금까지의 AI는 인간이 코드를 짜고, 데이터를 정제하여 학습시키는 개방형 루프였다. 발전의 속도는 인간의 노동력과 인지 속도에 종속되어 있었다. 그러나 2026년의 AI는 스스로 자신의 코드를 분석하고, 더 나은 구조로 고치며, 학습 방식을 최적화하는 폐쇄 루프를 형성한다. 이 루프가 닫히는 순간, 기술의 발전 속도는 산술급수(Linear)에서 기하급수(Exponential)로 전환될 소지가 커진다.

이와 관련해 AI의 지능 폭발(Intelligence Explosion)이라는 것은, 아이가 혼자 공부하는 법을 터득하면 부모의 도움 없이도 성적이 오르는 것과 같다. AI가 더 똑똑해질수록, 자신을 더 똑똑하게 만드는 방법 또한 스스로 더 잘 찾아내게 된다. 한 번 시작되면 기하급수적으로 똑똑해지는 이 과정은 멈출 수 없는 가속 페달과 같으며, 이것이 바로 Anthropic이 경고하고 대비하는 폭발적 가속의 실체다. 이는 인간이 실시간으로 추적하거나 이해할 수 없는 영역으로 진입함을 뜻한다. 이해하기가 벅찬 와중에 계속해서 뛰어난 무언가가 세상에 변화를 주고 있다는 뉴스가 쏟아지니 AI에 관심 있는 모든 이들이 피로감을 느끼는 것이다.

(2) Frontier Red Team: 안전에서 통제로

이 기술적 특이점은 안전의 정의를 송두리째 바꾼다. Anthropic의 Red Team 채용 공고와 Dario Amodei의 지난 1월 에세이 "The Adolescence of Technology"는 이러한 변화를 명확히 보여준다.

AI 모델을 만드는 회사라면 모두 가지고 있는 Red Team의 역할은 원래, "AI가 욕설이나 차별적 발언을 하는가?"를 감시하는 수동적 관리자라고 볼 수 있었다. 그러나 2026년의 Anthropic에서 정의하는 Frontier Red Team은 "스스로 업그레이드하는 AI를 인간이 어떻게 통제할 것인가?"를 연구하는 능동적 방어군이다. Anthropic의 전략은 그런 위험한 AI를 우리가 먼저 만들어보고, 그 위험성을 파악한 뒤 방어법을 찾자는 것이다. 외부에서 통제 불가능한 초지능이 등장하기 전에, 내부의 통제된 환경에서 지능 폭발을 시뮬레이션하고 억제 메커니즘(Kill Switch)을 확보하려는 시도다. 이 회사의 시니어 개발자인 Sholto Douglas가 이미 테스트 중이라고 언급한 연속 학습 시스템(Continuous Learning Systems)은 바로 이 실험이 본격화되었음을 알리는 신호탄이다.

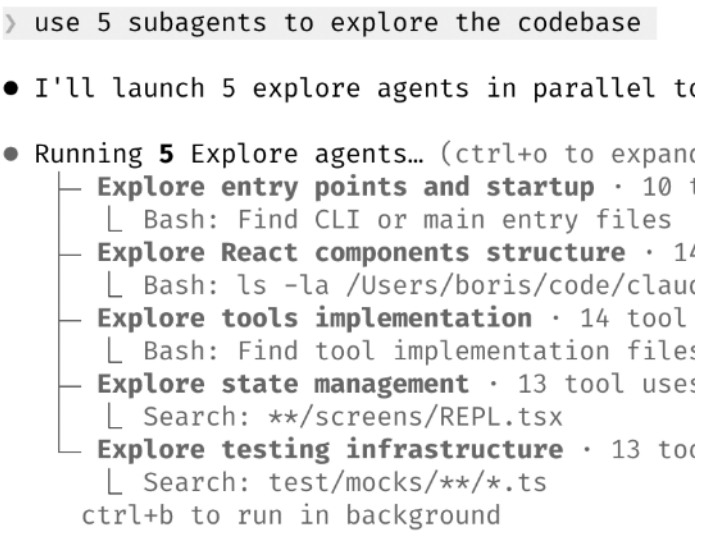
2. 충격적인 Claude Code 팀 업무 방식

(1) 작성의 시대에서 조율의 시대로

Node.js의 창시자 Ryan Dahl이 "구문(syntax)을 직접 작성하는 것은 더 이상 개발자의 일이 아니다"라고 선언한 것은 AI 시대를 맞은 개발자의 감상 젖은 레토릭으로만 치부할 것이 아니다. 실제로도 그렇다. Anthropic의 제품 핵심이라 할 수 있는 "Claude Code" 팀의 내부 작업 방식을 보면, 이미 인간이 코드를 한 줄 한 줄 짜는 디지털 수도승의 시대가 끝났음을 알 수 있다. 이제 개발자의 정의는 코더(Coder)가 아니라, 똑똑한 AI 비서 5명에게 일을 나눠주고 관리하는 팀장이자 지능의 조율자(Orchestrator)로 이동했다.

인간 개발자는 한 번에 하나의 생각만 할 수 있지만, AI는 무한히 분열할 수 있다. 예를 들어, 개발자는 5개의 터미널을 띄워놓고, 터미널 1에서는 프론트엔드 리팩토링, 터미널 2에서는 백엔드 API 수정, 터미널 3에서는 로그 분석이 동시에 돌아가게 할 수 있다. 쉽게 말해, 5명의 에이전트에게 동시에 별도의 일을 시키고, 본체인 인간은 커피 마시면서 결과만 확인하면 되는 것이다. 이러한 일처리 방식은 마치 CPU 멀티코어 프로세싱과 같다. 인간 개발자는 CPU의 스케줄러가 되어 5개의 Claude Agent에게 작업을 분산시키고 가장 상위의 결과만 취합하는 일처리 방식이다.

그림 3. 인간이 "5개의 서브 에이전트를 써라"라고 자연어로 명령하자,
메인 에이전트가 즉시 하위 인스턴스를 복제해 프론트엔드 탐색부터 테스트 구조 파악까지
5가지 업무를 동시에 '병렬 처리(Parallelism)'하는 모습.



자료: Anthropic, 미래에셋증권 리서치센터

나아가, Claude Code를 창시한 Boris Cherny는 6개월 넘는 기간 동안 SQL(Structured Query Language)조차 한 줄도 안썼다고 고백했다. "지난주 지표 봅아줘"라고 말하면 Claude가 알아서 SQL 코드를 짜고 실행해서 결과만 보여주기 때문이다. 데이터베이스는 기본적으로 SQL이라는 매우 엄격한 문법의 언어로만 대화가 통했는데, 이제는 인간은 자연어만 쓰고 AI가 통역사가 되어 기계어로 알아서 바꿔주는 세상이다. 우리들이 배워야 할 언어는 영어, 단 한 가지일지도 모르겠다.

(2) 재귀적 자기 수정과 집단지성

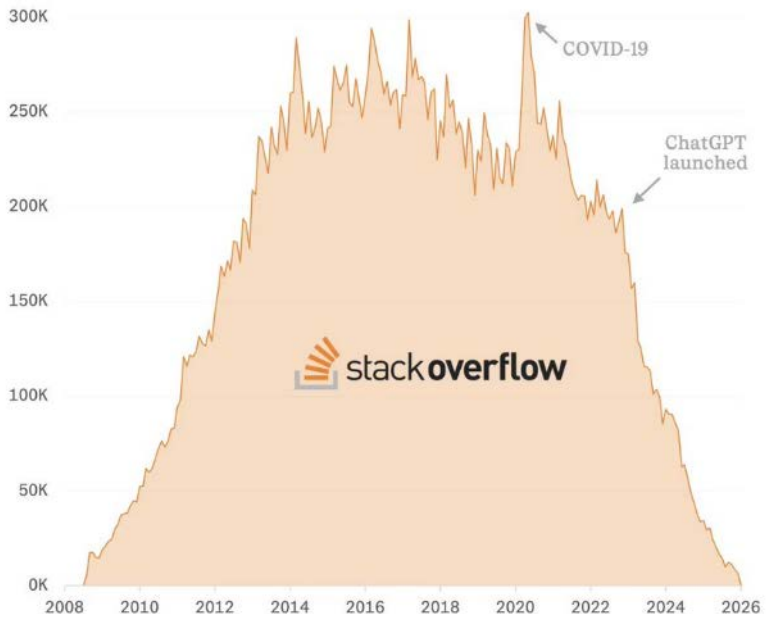
Claude Code의 가장 강력한 점은 코드를 생성하는 것이 아니라, 스스로 검증하고 수정하는 재귀적 루프에 있다. 복잡한 작업일수록 바로 코딩에 들어가지 않고, Plan Mode에서 설계를 먼저 수행한다는 점이 특이하다. 예를 들어, 한 Claude 에이전트가 계획을 짜면, 다른 에이전트(Staff Engineer 페르소나)을 띄워 그 계획을 리뷰하게 한다. "AI가 짜고 AI가 감수하는" 구조다. 계획이 완벽해지면, 코딩 구현은 문맥 전환 없이 한 번에(One-Shot) 이루어진다.

Anthropic의 가장 하이엔드 모델인 Claude Opus 4.5의 지능을 탑재한 Claude Code는 단일 모델이 아닌 에이전트 군단(Swarm)으로 작동한다. 메인 에이전트이자 팀장 역할을 맡은 Claude가 인간 사용자의 명령을 받으면, 스스로 판단하에 새로운 Claude 인스턴스(부하 직원)를 API로 호출한다. Master Agent가 작업을 조율하면, Sub-agent 1은 파일 탐색, Sub-agent 2는 구조 파악, Sub-agent 3는 테스트 코드 작성을 수행하는 식이다. 그리고 서브 에이전트들은 본인의 업무를 수행한 뒤 요약본만 팀장 에이전트에게 송부한다. 그야말로 재귀적 에이전트 호출(Recursive Agent Calls)이다.

Claude Code 팀은 "병렬 처리(Parallelism), 에이전트 위임(Delegation), 그리고 재귀적 자기 수정(Recursive Self-Correction)"을 통해, 인간 1명이 수행하던 업무량을 10배 이상 확장했다고 밝혔다.

과거의 병목이 코드를 얼마나 빨리 치느냐였다면, 이제 병목 구간은 코드 작성이 아니라 인간의 리뷰 속도이다. AI가 쏟아내는 방대한 양의 고품질 코드를 인간이 검증하고 승인하는 속도가 전체 개발 속도를 결정하게 되었다. 이는 개발자가 작가가 아니라 편집장이 되어야 함을 시사한다.

그림 4. Stack Overflow의 질문 수가 ChatGPT 출시 이후 급격히 붕괴하여 2026년에는 바닥
"인간이 코드를 짜고 질문하던 시대의 종말"을 시각적으로 선언
개발자 생태계의 지식 전수 방식이 '집단지성'에서 '인공지능'으로 완전히 대체



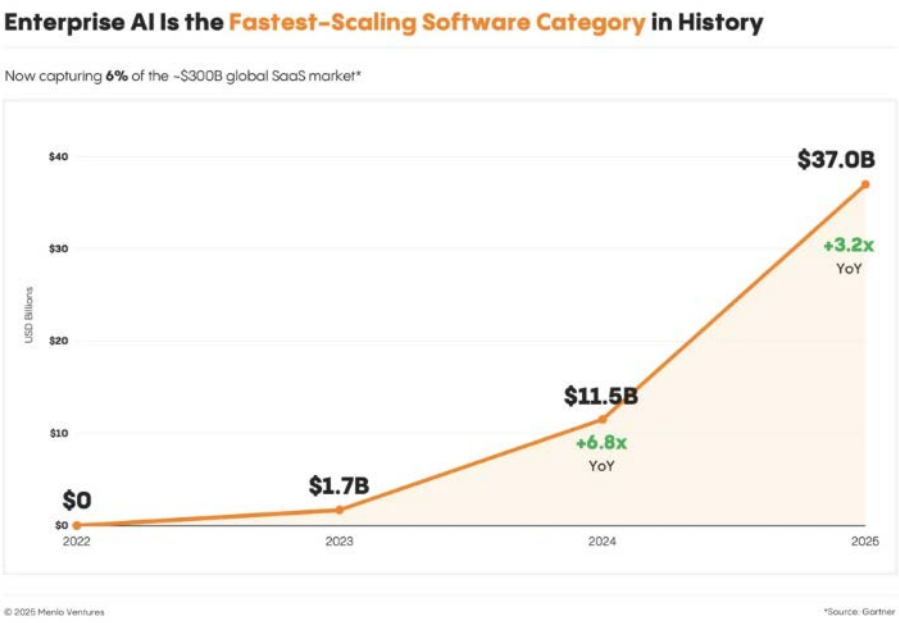
자료: Stack Overflow, 미래에셋증권 리서치센터

3. AI 엔터프라이즈 시장은 구조적 붐을 가리킨다

(1) PoC(개념 증명)에서 프로덕션으로의 자금 이동

Menlo Ventures의 2025년 보고서는 AI 시장을 둘러싼 거품론에 대한 좋은 반박 자료다. Menlo 측의 설문에 따르면, 2024년까지 기업들의 AI 관련 지출은 대부분 R&D나 별도로 책정된 혁신 실험비(Innovation Budget)에서 집행되었다. 이는 "한번 써보고 아니면 말자"는 식의 가벼운 접근이었다. 그러나 2025년 기록된 $37B의 지출은 대부분 기업의 핵심 IT 인프라 및 고정적인 운영 비용(OpEx) 계정에서 나왔다.

그림 5. 엔터프라이즈 AI는 2024년 115억 달러에서 2025년 370억 달러로 3.2배 폭증
닷컴 버블이나 모바일 전환기보다 훨씬 빠른 속도로 AI가 실물 경제의 손익계산서에 침투
전체 SaaS 시장의 6%를 단숨에 점유한 이 수치는, AI 도입이 필수재(Must-have)임을 시사



자료: Menlo Ventures, 미래에셋증권 리서치센터

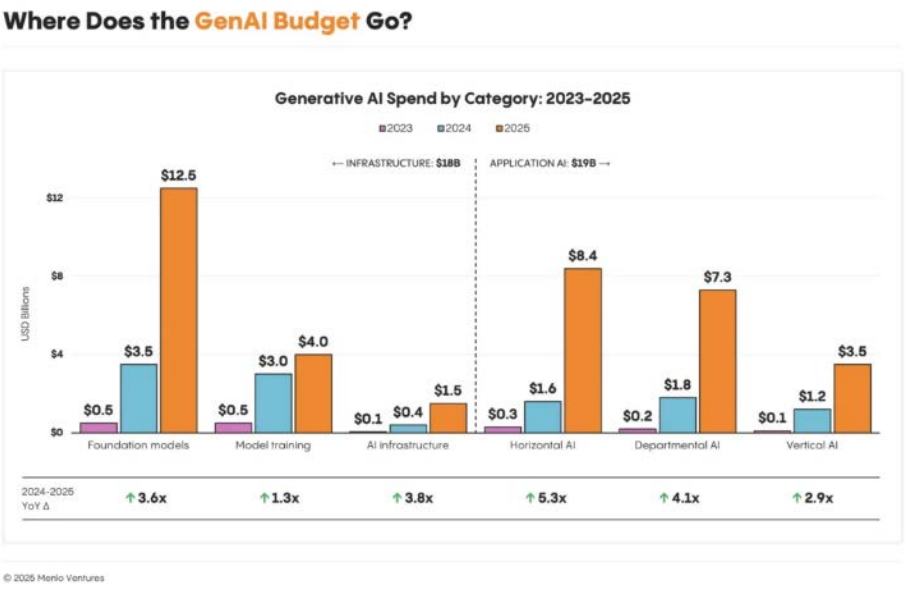
이는 AI가 기업 경영의 선택 사항이 아닌, 전기나 클라우드 서버와 같은 필수 인프라로 격상되었음을 의미한다. 특히, 생산 단계 적용률(Production Rate)이 47%에 달한다는 점은 AI가 기업의 핵심 워크플로우에 깊숙이 통합되었음을 증명한다.

그리고 중대한 넘버는 엔터프라이즈 생성 AI 지출이 2024년 $11.5B에서 2025년 $37B로 1년 만에 3.2배 폭증했다는 사실이다. 이는 엔터프라이즈 소프트웨어 역사상 가장 가파른 성장 속도이며, 클라우드 컴퓨팅이나 모바일 전환 당시보다 훨씬 빠른 카테고리 확장을 보여준다.

또한, 닷컴 버블 당시와 현재 AI 붐의 가장 큰 차이점은 인프라와 애플리케이션의 균형에 있다. 닷컴 버블 때는 광케이블과 서버(인프라) 투자는 활황을 이뤘으나, 그 위에서 돈을 버는 서비스(애플리케이션)는 부재했다. 그러나 2025년 데이터는 인프라($18B)와 애플리케이션($19B) 투자가 거의 50:50으로 완벽한 균형을 이루고 있음을 보여준다. 기업들이 엔비디아의 GPU(인프라)를 사는 이유는 유행을 좇는 게 아니라, 그 칩 위에서 돌아가는 애플

리케이션이 실제 수익을 창출하고 있기 때문이다. 칩을 사서 모델을 돌리고, 그 모델이 실제로 돈을 벌어다 주거나 비용을 획기적으로 절감해 주는 선순환 구조가 완성되고 있는 것이다. 이는 시장이 장기적인 성장을 위한 탄탄한 펀더멘털을 갖춰가고 있음을 시사한다.

그림 6. 인프라와 애플리케이션 투자가 거의 50:50으로 완벽한 균형
닷컴 버블 당시 인프라에만 돈이 쏠렸던 기형적 구조와 달리, 건전한 수익 모델 위에 서 있음. AI 기반 애플리케이션이 실질적인 수익(ROI)과 비용 절감을 만들어내는 선순환 구조의 방증.

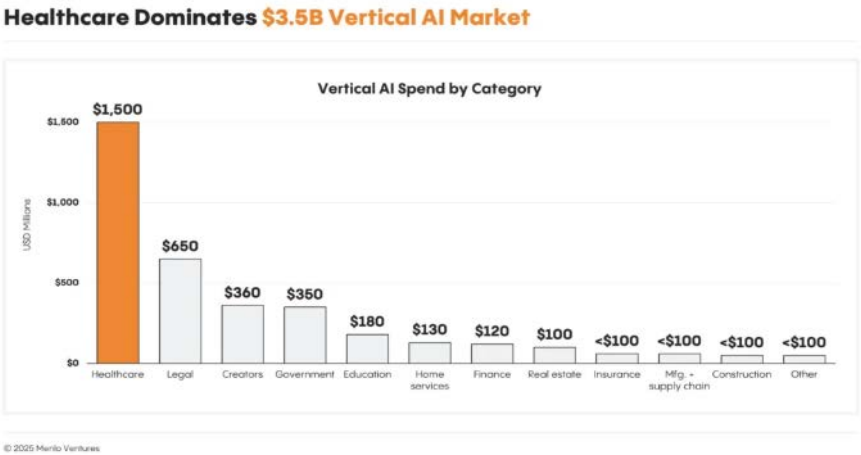


자료: Menlo Ventures, 미래에셋증권 리서치센터

(2) 킬러 앱의 등장과 옥석 가리기

AI 애플리케이션 시장의 성장은 모호한 생산성 도구가 아닌, 명확한 ROI(투자 대비 수익)가 나오는 분야에서 주도되고 있다. 특히 주목할 분야는 코딩 어시스턴트 시장($4B)과 헬스케어(Scribes 분야, $600M) 시장의 폭발적 성장이다.

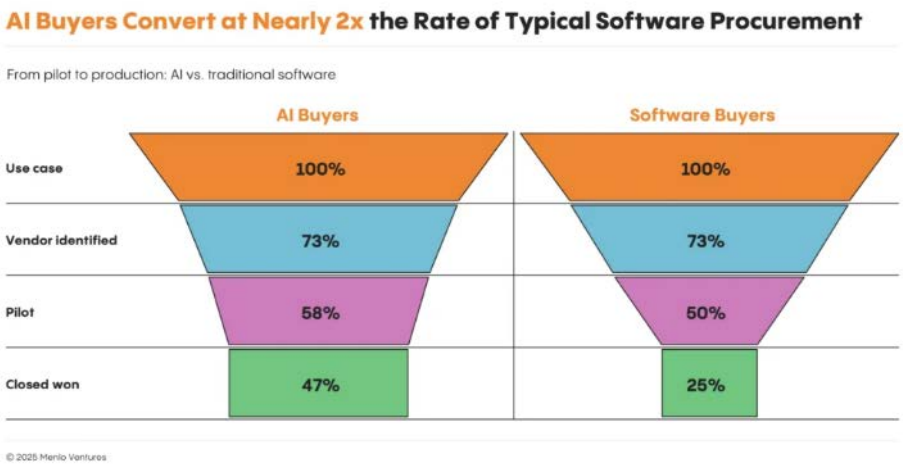
그림 7. 헬스케어 분야가 15억 달러로 압도적 1위를 차지
의사의 진료 기록(Scribe) 업무처럼 고임금 노동 시간을 줄여 인건비 절감 효과가 확실한 곳에서부터 돈이 몰리고 있다는 방증. 헬스케어처럼 전문 지식이 깊은 영역일수록 AI는 필수재로 자리 잡는 중.



자료: Menlo Ventures, 미래에셋증권 리서치센터

이 두 영역은 AI가 인간의 노동 시간을 물리적으로, 그리고 즉각적으로 대체할 수 있는 곳이다. 개발자가 코드를 짜는 시간, 의사가 진료 기록을 작성하는 시간을 AI가 대신함으로써 기업은 즉각적인 인건비 절감 효과를 누린다. 이런 킬러 앱들에 힘입어, AI 소프트웨어의 유료 전환율은 기존 SaaS의 2배에 달한다. 사용자들이 AI 기능을 있으면 좋은(Nice-to-have) 기능이 아니라, 없으면 업무가 안 되는(Must-have) 필수재로 점차 인식하고 있음을 보여준다.

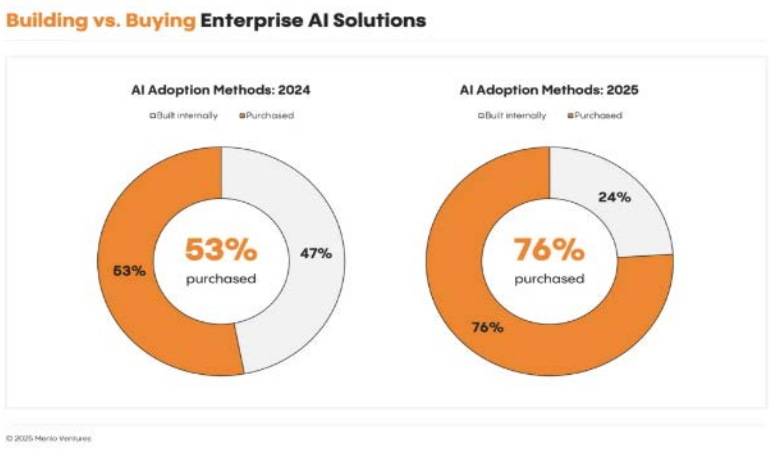
그림 8. AI 솔루션의 파일럿 대비 구매 전환율이 47%로 일반 소프트웨어(25%)의 약 2배에 달한다
과거의 SaaS가 긴 설득 과정을 거쳐야 했다면, AI는 즉각적인 생산성 향상과 비용 절감 효과를 증명
기업 고객의 니즈가 기능(Feature) 검토에서 실질적 결과(Outcome) 확인으로 이동했음을 보여줌



자료: Menlo Ventures, 미래에셋증권 리서치센터

한편, 기업들의 AI 도입 전략은 "Buy vs Build(살 것이냐 만들 것이냐)"라는 이분법적 논쟁을 넘어, "Buy to Build(사서 구축한다)"라는 실용적 하이브리드 모델로 정착한 것으로 풀이된다. 자체적으로 LLM을 밑바닥부터 학습(Pre-training)시키는 무모한 도전을 하는 기업은 이제 별로 없다. 기업의 76%가 모델을 직접 훈련하는 대신 Anthropic, OpenAI, Google 등의 외부 상용 모델을 구매(Buy)한다. 이들의 성능 발전 속도를 개별 기업이 따라잡는 것은 불가능하기 때문이다.

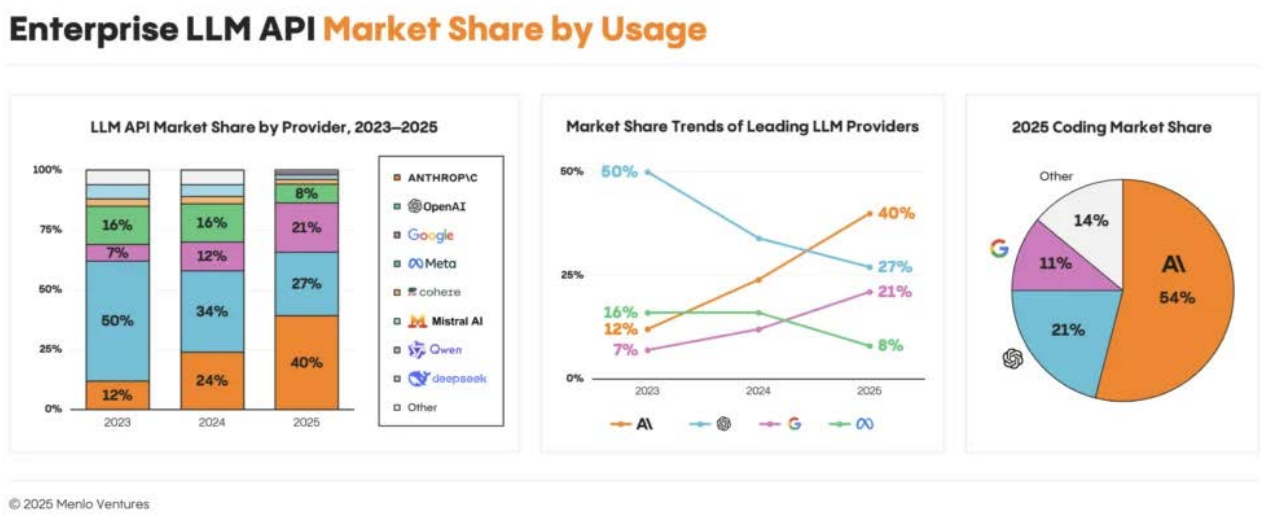
그림 9. 자체 모델 개발 비율이 급감하고 외부 상용 모델 구매 비율이 76%로 압도적인 비중 차지



자료: Menlo Ventures, 미래에셋증권 리서치센터

대신 기업들은 구매한 모델을 그대로 쓰지 않는다. 자사의 데이터와 결합하고, 자사의 워크플로우에 맞게 애플리케이션 로직을 직접 구축(Build)하는 데 역량을 집중한다. 모델은 외부에서 빌려오되, 그 모델이 일하는 방식은 내부에서 통제하는 것이다. 이 과정에서 엔터프라이즈 시장의 패권은 Anthropic으로 기울고 있다. 기업 고객들은 모델의 재미나 창의성보다 데이터 보안, 긴 컨텍스트 윈도우(Context Window), 그리고 복잡한 업무 지시를 정확히 수행할 수 있는 에이전트 능력을 최우선 기준으로 삼고 있다. Anthropic의 Claude는 이러한 기업의 니즈를 가장 잘 충족시키며 엔터프라이즈 시장 점유율 1위를 차지했다.

그림 10. 2023년 12%에 불과했던 Anthropic의 점유율이 2025년 40%로 치솟으며 OpenAI를 제치고 1위 유지
특히 코딩 시장에서 Anthropic이 54%라는 과반 점유율을 차지한 데이터는 '에이전트 능력'이 출중한 모델에 지갑을 열고 있다는 증거



자료: Menlo Ventures, 미래에셋증권 리서치센터

시장이 성숙해지면서 "가짜 AI 기업들"에 대한 구조조정도 가속화되고 있다. GPT API를 단순히 포장만 해서 내놓았던 얇은 서비스(Wrapper)들은 멸종 위기를 맞고 있다. 거대 모델사들이 기능을 내재화(예: Claude Cowork, ChatGPT Search)하면서 설 자리를 잃었기 때문이다.

반면, 연간 반복 매출(ARR) $100M을 돌파한 AI 제품이 50개를 넘어섰다. 이들은 독자적인 워크플로우와 데이터 처리 기술을 통해 고객에게 실질적인 가치를 제공하는 진짜 제품들이다. 이 50개의 신규 제품들은 기존 레거시 SaaS의 파이를 잠식하며 성장한 것이며, 이는 AI 시장이 기대감의 단계를 넘어 실적의 단계로 완전히 진입했음을 증명하는 증거다.

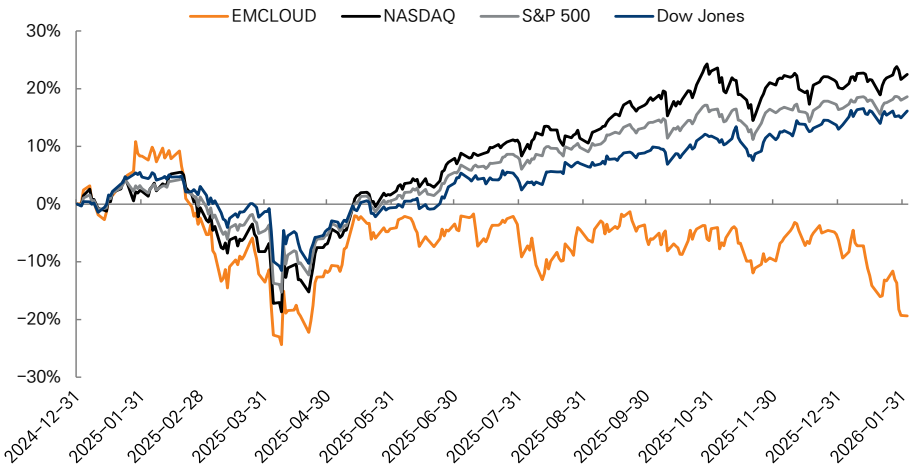
4. Anthropic 쇼크와 SaaS의 위기

(1) 밸류에이션의 붕괴: 예측 가능성의 상실

* EMCloud 지수: 주요 클라우드·소프트웨어 기업을 포함하는 지수

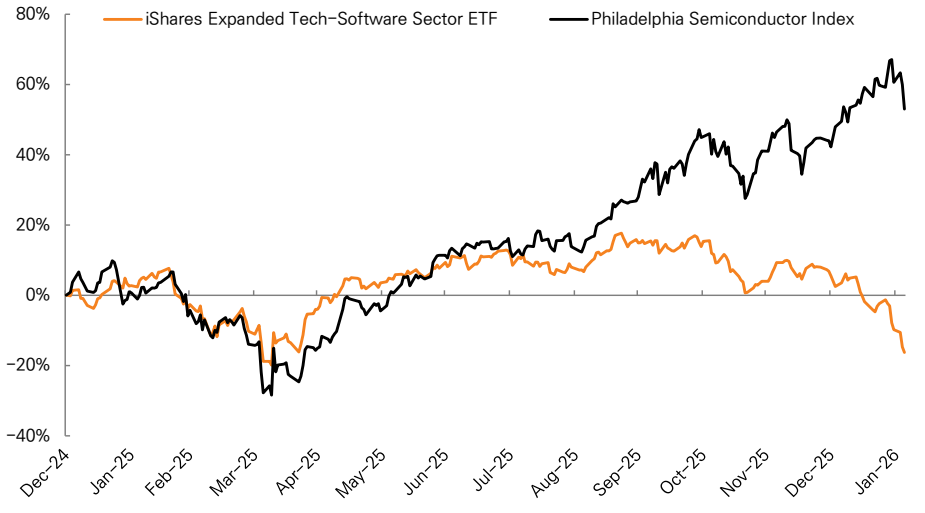
빛이 있으면 그림자가 있다. AI 엔터프라이즈 시장의 고공행진과 달리, 지난 15년간 테크 산업을 지배해 온 SaaS(Software as a Service) 모델의 신화는 무너지고 있다. S&P 500이 사상 최고치를 경신하는 강세장 속에서도, 클라우드 소프트웨어 지수(EMCLOUD: BVP Nasdaq Emerging Cloud Index)가 31.4% 폭락했다는 사실은 자본 시장의 냉혹한 재평가를 보여준다.

그림 11. EMCLOUD와 주요 지수 비교
AI 시대에 소프트웨어 기업들의 성장 둔화 압력이 반영



자료: Bloomberg, 미래에셋증권 리서치센터

그림 12. iShares Expanded Tech-Software ETF vs Philadelphia Semiconductor Index
반도체와 소프트웨어 주식 사이에 커다란 퍼포먼스 차이가 벌어짐
현재, 시장에서 AI 관련 수혜가 어디에 쏠렸는지 보여주는 적나라한 그래프

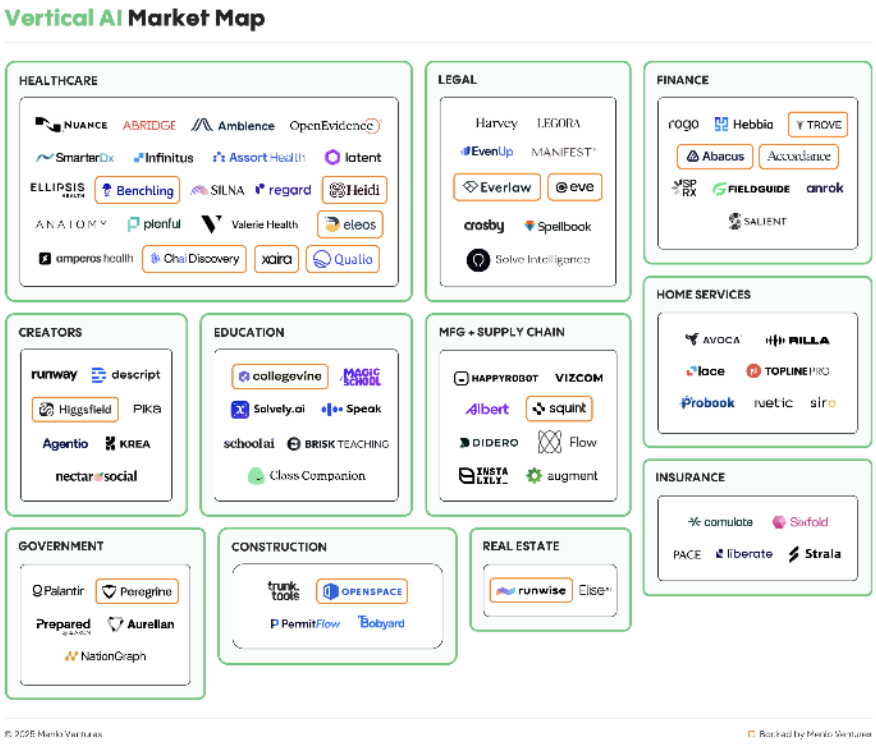


자료: Bloomberg, 미래에셋증권 리서치센터

*** 버티컬(Vertical) 소프트웨어:** 특정 직업이나 산업에 종사하는 사람들만 쓰도록, 그 일에 딱 맞춰서 만든 전문적인 소프트웨어

과거 SaaS 기업들은 적자 상태에서도 높은 PER(주가수익비율)을 인정받았다. 그 근거는 높은 이탈 비용(=Lock-in effect)과 반복 매출(ARR)의 예측 가능성이었다. 그러나 AI의 등장은 이 두 가지 전제를 모두 파괴하는 것처럼 보인다. 특히, 데이터 제공업체(Data-Provider) 및 소프트웨어 기업들의 주가 폭락은 Anthropic이 Claude Cowork Assistant에 법률(Legal) 기능을 추가한다고 발표한 직후 발생한 것으로 풀이한다. 이는 상징적인 사건이다. 수십 년간 축적된 전문 지식(Domain Knowledge)으로 쌓아 올린 버티컬 SaaS의 철옹성이, 범용 AI 모델의 기능 업데이트 하나로 무너질 수 있음을 시장이 인지하는 듯 보이기 때문이다. 자본 시장은 이미 소프트웨어 기업의 미래 현금 흐름을 굉장히 비관적으로, 그리고 회의적으로 재평가하기 시작했다.

그림 13. 현재 기준 Vertical 소프트웨어에서 AI-native 기업들 리스트



자료: Menlo Ventrues, 미래에셋증권 리서치센터

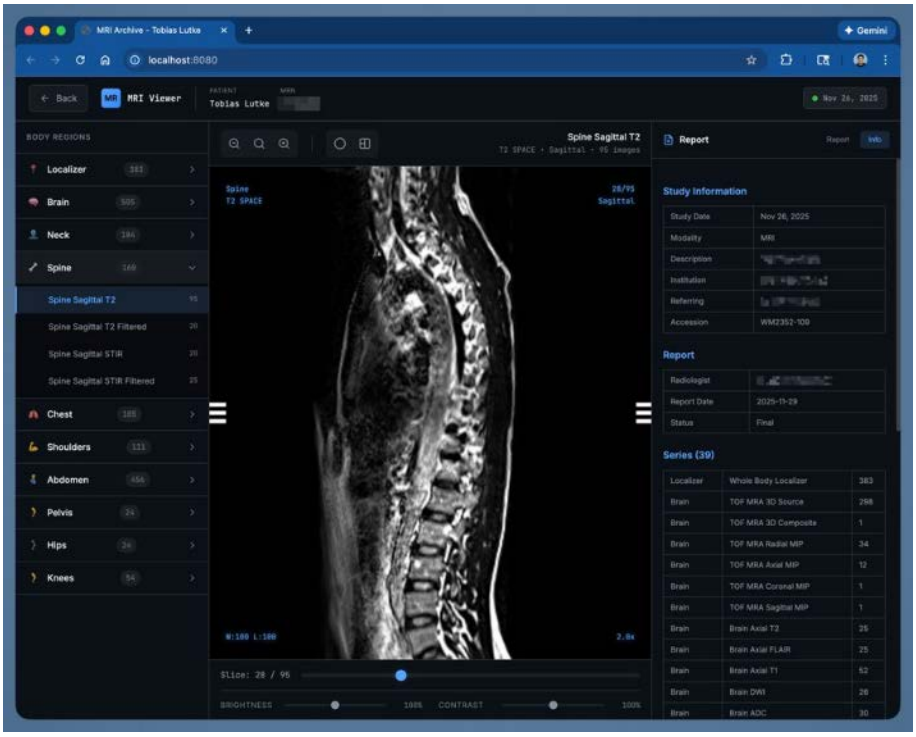
경제학의 기본 원칙에 따라, 공급이 무한대가 되면 가격은 0에 수렴한다. AI는 코딩 비용을 0으로 만들었고, 이는 소프트웨어의 공급을 폭발시켰다. 그런데 AI로 인한 코딩 비용의 하락은 양날의 검이다. 소프트웨어 기업은 더 효율적으로 코드를 짤 수 있지만, 이는 모든 경쟁자에게도 동일하게 적용되기 때문이다. 과거에는 소프트웨어를 만드는 것 자체가 진입 장벽이었으나, 이제는 누구나 만들 수 있다. 예를 들어, 어떤 회사가 HR 소프트웨어를 만든다면, 경쟁사 혹은 빅테크의 회계 소프트웨어가 하룻밤 사이에 HR 기능을 추가해 저렴한 가격으로 배포할 수 있게 됐다. 시장은 더 이상 새로운 파이를 만드는 게임이 아니라, 남의 파이를 뺏는 제로섬 게임으로 전환된 것이 아니냐는 의심을 갖고 있다. 모든 소프트웨어 기업이 서로의 영역을 침범하며, 결국 모든 기능이 범용화(Commoditized)되는 아찔한 레이스가 시작됐다는 논리다.

(2) 커스텀 솔루션의 역습

"Build vs Buy"의 경제학이 뒤집혔다. 과거에는 직접 만드는 것(Build)이 너무 비싸서 샀지만(Buy), 이제는 AI가 순식간에 만들어준다. 연간 약 5억 원을 CRM 특화 SaaS 업체에 지불하던 고객사들이 계약을 해지하고 자체 솔루션으로 전환하는 사례가 속출하고 있다는 이야기가 들린다. 고객들은 "왜 우리가 쓰지도 않는 수백 가지 기능이 덕지덕지 붙은 비싼 소프트웨어를 써야 하죠? AI한테 우리 회사에 딱 맞는 심플한 도구를 만들어달라고 하면 되는데요."라고 말하는 것이다.

이와 관련해, Shopify의 CEO Tobi Lutke는 고가의 의료용 윈도우 소프트웨어 대신 Claude를 시켜 즉석에서 MRI 뷰어를 만든 본인의 경험담을 공유했다. 이는 기존 소프트웨어의 복잡성과 라이선스 비용이 AI의 즉각적인 커스텀 코딩 능력에 의해 무력화되고 있다는 것을 시사한다. 필요할 때 생성해서 쓰고 버리는 소프트웨어 소비 패턴은, 고가의 라이선스 비용을 청구하는 전문 Vertical SaaS 기업들에게 치명적인 위협이다.

그림 14. Shopify CEO Tobi Lutke가 MRI 데이터를 보기 위해 Claude를 활용한 사례
복잡하고 비싼 전문의료 소프트웨어 대신, AI가 단 몇 초 만에 HTML 기반 커스텀 뷰어를 만들어냄.
사용자가 필요할 때 즉석에서 소프트웨어를 생성해서 쓰고 버리는 시대가 온 것.
고가의 라이선스 비용을 받는 버티컬 소프트웨어 기업들에게 치명적인 위협.



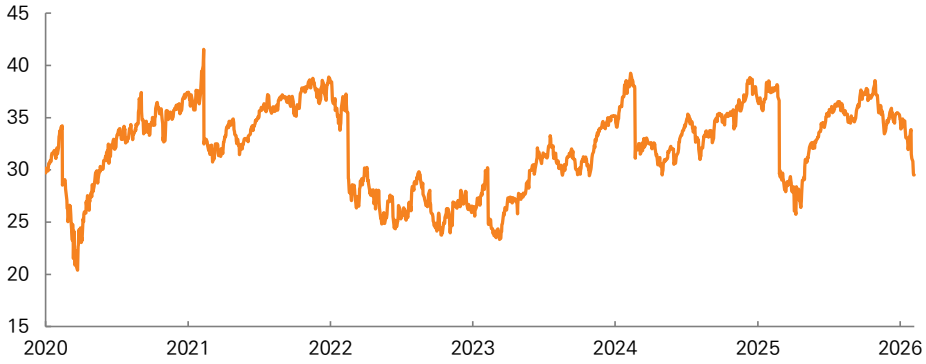
자료: Tobi Lutke, 미래에셋증권 리서치센터

MS의 CEO 사티아 나델라의 말대로, 소프트웨어 앱은 뒤쳐지고 있다. 과거의 소프트웨어는 그 자체로 똑똑했다. 복잡한 계산 로직과 업무 규칙이 앱 안에 내장되어 있었다. 그러나 이제 비즈니스 로직과 지능은 애플리케이션에서 AI 에이전트 레이어로 이동하고 있다. 사용자는 복잡한 CRM/ERP 화면을 직접 클릭하지 않고, 대신 "이번 분기 매출 예측 보고서 작성해"라고 에이전트에게 명령하는 세상이 된다. 에이전트가 백그라운드에서 DB에 접속해 작업을 수행한다. 그 과정에서 소프트웨어의 화려한 UI의 가치는 높게 쳐줄 수가 없다.

5. SaaS 종말론에 대한 반론

앞선 장들에서 우리는 AI가 가져올 파괴적 변화와 SaaS의 위기를 다루었다. 그러나 시장의 공포는 언제나 과장될 때가 많다. 2026년 현재, "SaaS는 끝났다"는 선언은 너무 성급하다고도 볼 수 있다. 이는 엔터프라이즈 소프트웨어의 본질을 기능(Feature)으로만 오독한 결과라고 생각한다.

그림 15. S&P 500 소프트웨어·서비스 그룹 PER 추이
현재는 30배 아래로 추락한 상황으로써 업종 전반으로 밸류에이션 압축이 진행되고 있음
대표적으로, Snowflake와 ServiceNow의 성장률은 20~30%대에서 횡보 중



자료: Bloomberg, 미래에셋증권 리서치센터

(1) 코드 비용 0이 제품 가치 0을 의미하지 않는다

시장은 "현재 AI가 코드를 공짜로 짠다. 그러니 소프트웨어 공급이 무한대가 된다. 그래서 SaaS 가격은 0이 된다"는 간단한 삼단논법에 빠져 있다. 그러나 이 등식에는 결함이 있다. 엔터프라이즈 세계에서 코드 작성은 전체 비용의 일부인 초기 계약금에 불과하다. 기업이 두려워하는 진짜 비용은 총소유비용(TCO), 즉 운영(Operation)에서 나온다.

코드는 AI가 이제 1초 만에 짤 수 있다. 그러나 보안 패치, 감사(Audit), 권한 관리(IAM), 장애 대응(On-call), 규정 준수(Compliance)가 붙는 순간, 그 소프트웨어는 코드만이 아니라 기업의 생명 유지 장치가 된다. AI가 코딩 비용을 0으로 만들어도, 이 운영 비용과 리스크 관리 비용은 0이 되지 않는다. 오히려 시스템이 빠르게 변할수록 운영의 난이도와 가치는 상승할 수 있다.

코드는 만들 수 있다. 그러나 문제는 굴리는 것이다. 굴리는 것은 만드는 것과 다른 차원의 노동이다. 기업이 소프트웨어 업체에 매년 돈을 지불하는 이유는 그들의 UI가 예뻐서가 아니다. 기업은 소프트웨어를 사는 것이 아니라, 비즈니스 프로세스 운영 대행(Business Process as a Service)을 사는 것이다. 기업은 기능이 아니라 신뢰성과 책임을 구매한다. 시스템이 멈췄을 때 멱살을 잡을 대상이 필요하고, 데이터가 유출되었을 때 책임을 질 계약 주체가 필요하다. 내부에서 AI로 뚝딱 만든 커스텀 툴은 이 '책임의 외주화'를 불가능하게 만든다.

게다가, "우리 회사에 딱 맞는 심플한 AI 툴"이 수십 개로 늘어나는 순간, 기업은 통합의 지옥을 맛보게 된다. 결국 이걸 다 통합해서 관리해 줄 "대장 소프트웨어"가 필요해진다. 즉, 데이터가 파편화되고 관리가 불가능해질수록, 역설적으로 모든 것을 통합하고 표준화해 주는 거대 플랫폼 SaaS의 지배력이 생겨날 여지도 있다. 다시 말해, AI가 강해질수록 조직은 소프트웨어를 더 빨리/더 자주 바꾸며, 그래서 운영·통합·거버넌스의 가치가 커질 수 있다는 말이다. 기업은 무한한 자유가 아닌, 통제 가능한 자유를 원하기 때문이다.

(2) 밸류에이션의 붕괴? 불확실성의 가격

현재 SaaS 기업들의 주가가 무너지는 현상은 비즈니스 모델의 사망 선고라기 보다는, 미래 현금 흐름의 불확실성이 생긴 것에 기인한다고 봐야 할 것 같다. 물론, 시장은 불확실성을 싫어하기 때문에, 불확실성에 대한 밸류에이션 멀티플 할인(Discount)은 불가피하다고 사료된다.

엄밀히 말하면, SaaS 시장은 과금 단위가 좌석(Seat)에서 사용량(Usage)이나 성과(Outcome)로 넘어가는 과도기를 겪고 있다고 볼 수 있다. 예전엔 사용하는 사람 수(Seat)대로 돈을 받았다면, 앞으론 AI가 얼마나 많은 일을 처리했는지나 어떤 결과를 냈는지로 돈을 벌게 된다는 말이다.

그리고, AI 에이전트가 사람을 대체하면 좌석 숫자는 줄어들겠지만, 에이전트가 사용하는 API 트래픽과 데이터 처리량은 폭증한다. 다시 말해, AI가 코드를 더 많이 뽑을수록, 오히려 관측가능성, 정책/권한, 테스트·릴리즈(DevOps), 데이터 품질 같은 운영 레이어 수요가 증가할 수 있다. 이건 전형적 SaaS 영역이다. 결과(outcome)를 보장하고 책임을 질 수 있는 "Deep SaaS(락인 효과가 거대한 SaaS)" 기업들에게 이 변화는 시장 축소가 아니라 오히려 거대한 TAM의 확장을 의미한다.

이와 관련해 에이전트가 제멋대로 날뛰지 않게 하려면 어떤 데이터에 접근 가능한지(권한), 어떤 행위가 허용되는지(정책), 무슨 일을 했는지(감사 로그)를 통제하는 중앙 시스템이 필수적이다. SaaS는 이제 사람이 클릭하는 화면에서 AI 에이전트를 통제하고 감시하는 중앙 사령부로 진화해야 하고, 또 그럴 것이다. 따라서 "SaaS는 이제 끝났다"는 과장된 공포라고 생각하고, 더 정확히 말하면 소프트웨어 업계에서도 확실한 양분이 일어날 것으로 전망한다.

- **위기의 존재**: AI 전환에 실패 중인 레가시. AI API를 포장했을 뿐인 Wrapper, 단순 기능(Feature) 단위의 툴, 전환 비용이 낮은 얇은 SaaS.

- **기회의 존재**: 복잡한 프로세스를 운영하는 플랫폼, 강력한 통합과 보안을 제공하는 시스템, 고객사의 핵심 현실 데이터의 원천을 처리하는 Deep SaaS.

단순한 LLM 래퍼 서비스나 챗봇 형태의 SaaS는 살아남지 못할 것이다. 누구나 LLM을 쓸 수 있는 시대에 차별화는 기업의 복잡한 내부 데이터를 어떻게 AI가 이해하고 행동하게 만드느냐에 달려 있기 때문이다. 반대로, 회사의 핵심 데이터(장부, 원천 정보)를 꽉 쥐고 있거나, 복잡한 업무를 빈틈없이 처리해 주는 Deep SaaS는 살아남아 더 비싸질 것이다.

결론적으로, AI는 소프트웨어의 단위를 도구에서 운영과 책임으로 바꾸는 것이다. 이러한 관점의 변화에서, 살아남은 승자는 더 강력한 해자를 갖게 될 것이며, 그들의 서비스는 더 비싸질 것이다. 승자는 “툴 판매자”가 아니라 “결과 판매자”가 된다.

6. 종말은 없다, 'Deep SaaS'의 탄생이 있을 뿐

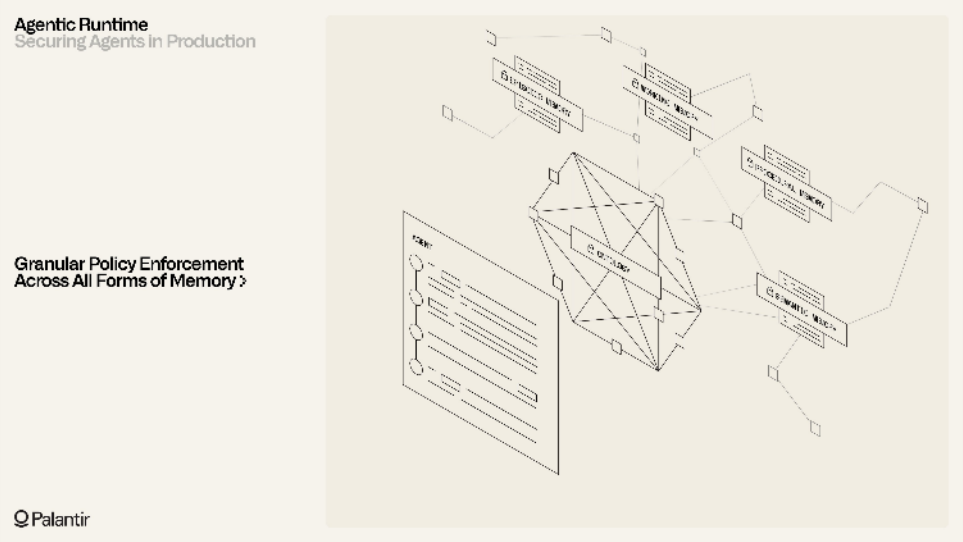
(1) Deep SaaS의 본질은 결국 온톨로지다

우리가 앞서 논의한 "AI 에이전트를 위한 중앙 통제 시스템, 권한 관리, 행위의 기록"이라는 개념은 사실, 팔란티어(Palantir)가 지난 20년간 구축해 온 온톨로지(Ontology)의 철학과 정확히 일치한다.

오히려 팔란티어는 AI 붐이 오기 전부터 '운영의 무결성'과 '결정론적 통제'를 준비해 온 기업이며, LLM 시대에 그 가치가 재발견되고 있는 가장 대표적인 사례라고 할 수 있다. LLM은 확률적(Probabilistic)으로 말을 지어내지만, 기업의 운영은 결정론적(Deterministic)이어야 한다. 팔란티어의 온톨로지는 기업의 데이터, 자산, 프로세스를 디지털 트윈으로 매핑하여 LLM을 현실 세계에 그라운딩(Grounding)시킨다. 즉, AI가 아무 말이나 하는 것을 막고, 정의된 온톨로지(현실의 규칙) 내에서만 사고하고 행동하게 강제하는 가드레일 역할을 한다.

온톨로지라는 것은 "에이전트를 위한 4차원 기억 저장소"로 생각하면 된다. 현재의 LLM은 기억 상실증이라는 한계를 가진다. 그러나 기업의 업무는 연속적이다. 팔란티어의 온톨로지는 단순 DB가 아니라, 에이전트에게 작업 기억(Working), 일화적 기억(Episodic), 의미적 기억(Semantic), 절차적 기억(Procedural)의 4가지 메모리를 통합 제공하는 두뇌의 해마 역할을 한다. 이는 앞으로의 Deep SaaS가 되기 위해서는, 에이전트가 과거의 성공 사례를 조회하고(Episodic), 회사의 고유한 맥락/지식을 학습하며(Semantic), 정해진 절차대로 행동하게 만드는(Procedural) 기업의 집단 지성 플랫폼으로 진화해야 함을 시사한다.

그림 16. 온톨로지가 에이전트에게 4가지 기억(일화적, 작업, 의미적, 절차적)을 제공하여, 기억이 없는 LLM을 '맥락을 이해하는' 직원으로 진화시키는 과정을 시각화
기억을 토대로, AI의 행동 하나하나가 회사의 정책과 과거 성공 사례에 기반하도록 검증을 수행
에이전트가 "왜 그런 판단을 했는지" 역추적 가능하게 만들어 기업이 에이전트를 안심하게 쓰게끔 함



자료: Palantir, 미래에셋증권 리서치센터

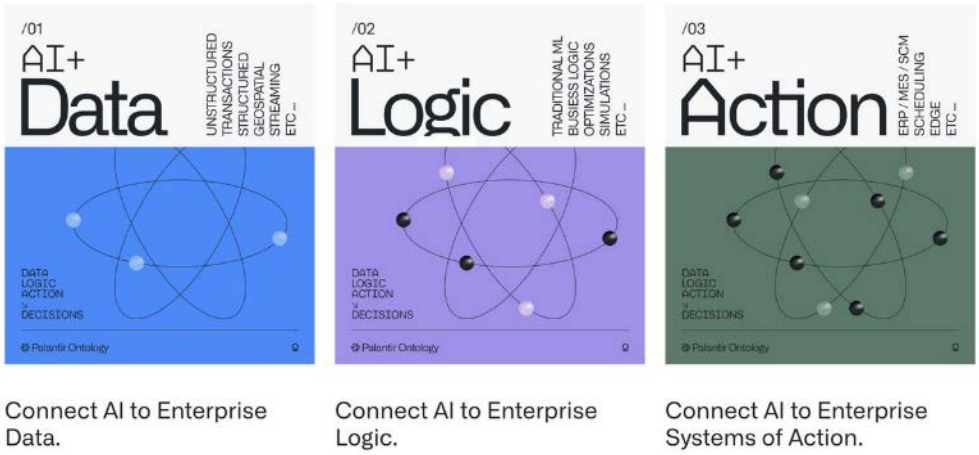
애초에 팔란티어의 CEO 알렉스 카프는 팔란티어가 단순한 도구가 아니라, 기업의 운영 체제(OS)가 되어야 한다고 주구장창 주장해온 인물이다. 기업이 AI를 통해 비용을 절감하거나 수익을 창출하는 구체적 결과를 만들어내지 못하면 그 소프트웨어는 가치가 0에 수렴한다는 것이 그의 지론이다. 심지어, 알렉스 카프는 기존 SaaS 비즈니스 모델을 "유령 소프트웨어(Shelfware)를 파는 세일즈 게임"이라고 비난하기도 했다. SaaS 기업들이 ARR(연간 반복 매출)을 늘리기 위해 고객에게 필요 이상의 기능(Feature)과 계정(Seat)을 강매한다는 것이다. 즉, 고객의 본질적인 문제를 해결하는 것이 아니라, 소프트웨어 접근 권한만 파는 행위라고 쏘아붙인 것이다. 그리고 앞서 우리가 언급한 대로, 이런 마케팅 업셀링에 기댄 회사가 AI 시대에 가장 큰 타격을 입을 가능성이 높다.

반면, 팔란티어의 엔지니어들은 직접 고객의 현장에 투입되어 고객의 데이터와 현실 세계를 연결하는 온톨로지를 구축한다. 이런 경영 철학은 2023년 말부터 본격화된 AIP(Artificial Intelligence Platform) 부트캠프 전략으로 구체화된다. AIP Bootcamp는 “Show, don’t tell”의 구현이다. 짧은 기간(약 일주일)에 고객 데이터로 실제로 작동하는 유즈케이스를 만들어 성과를 먼저 보여주는 방식이다. "돈부터 내라"가 아니라 "결과를 봐라, 가치가 있으면 사라"는 식이고, 이는 SaaS 업계의 관행(장기 계약 후 설치)을 뒤집은 것이다.

(2) 팔란티어는 그냥 SaaS가 아니라 왜 AI 시대의 운영체제

AI 에이전트 시대가 찾아옴에 따라, 각 회사들은 각자의 필요에 의해 다수의 소프트웨어를 직접 Build하게 된다. AI로 회사 내부에 자잘한 툴을 수십 개 만들면, 파편화된 툴의 지옥이 올 수 있다고 우리는 앞서 밝혔다. 그리고 이때 당연히 중앙의 컨트롤 타워가 있어야 하는데, 팔란티어가 가진 "AIP의 온톨로지 통합"이 바로 그 역할을 한다. 특히, AIP Logic과 Action으로 만들어지는, 수많은 마이크로 앱들은 파편화되지 않는다. 왜냐하면 그 뿌리가 모두 하나의 온톨로지에 연결되어 있기 때문이다.

그림 17. AI + Data / Logic / Action (Ontology Pillars)
3단계 구성은 LLM을 기업 환경에 '그라운딩' 시키기 위한 팔란티어 온톨로지의 3대 요소
AI가 현실 세계의 '물리적 진실(System of Truth)' 위에서만 사고하도록 강제
Logic은 기업 고유의 비즈니스 규칙을 AI에 주입하여 환각을 억제,
Action 파트는 ERP/MES 시스템에 실제로 기록하고 발주를 넣는 행위를 하게끔 함.
범용 모델(OpenAI, Google)이 쉽게 침범할 수 없는 영역, 기업의 로직이 쌓일수록 대체 어려움



자료: Palantir, 미래에셋증권 리서치센터

즉, 팔란티어는 AI 시대의 무질서(파편화, 환각, 보안 사고)를 질서(온톨로지, 가드레일, 최고 등급 보안)로 바꿔주는 플랫폼이다. 이것이 바로 그들이 단순한 툴이 아니라 기업의 중추신경계가 될 수 있는 이유다.

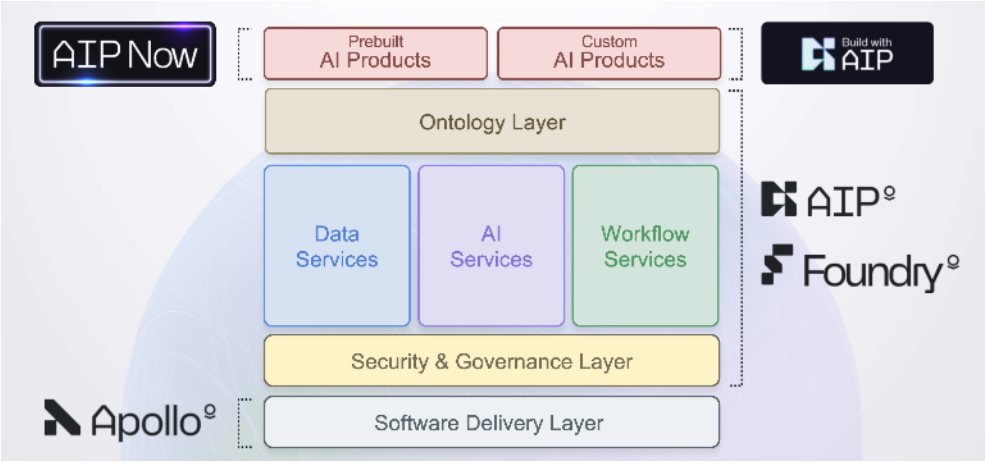
표 2. AI로 자체 구축하면 파편화·통제·신뢰의 3대 문제가 생긴다: 팔란티어가 'Deep SaaS'가 되는 이유

AI 시대의 소프트웨어의 핵심 문제	일반 SaaS 혹은 내부적으로 AI로 SW 만들 때의 한계	팔란티어 AIP의 솔루션	시장 포지셔닝
파편화된 툴의 지옥	각 툴이 독립된 DB 사용 → 데이터 사일로 발생 → 통합 불가능	온톨로지: 모든 마이크로 앱이 단일 진실의 원천에 연결 → AIP Logic으로 수백 개 툴을 만들어도 자동 통합	기업의 운영 체제 (OS)
제멋대로인 AI 통제 불가	AI 모델의 환각+ 권한 관리 부재 → 보안 사고 위험	AIP Control Plane + 가드레일: AI 행동 전후 권한 검증 + Human-in-the-loop 강제 → 통제 가능한 자율성	AI 중앙 사령부
Deep SaaS로서의 해자 부족	얕은 보안 인증 + 낮은 신뢰성 → 미션 크리티컬 업무에 부적합	FedRAMP High + DoD IL6: 미 정부로부터 기밀급 보안 인증 → 대체 불가능한 신뢰 자산	미션 크리티컬 Deep SaaS

자료: 미래에셋증권 리서치센터

일반적인 사내 툴은 각자의 DB를 따로 쓰거나 엑셀 파일로 돌아다녀 데이터 사일로(Silo)를 만들지만, 팔란티어의 온톨로지는 회사의 모든 실물 자산(공장, 트럭, 재고, 고객 등)과 그 관계를 정의한 단 하나의 진실(Single Source of Truth)이다. 개발자가 AIP Logic으로 100개의 다른 앱을 만들어도, 이 앱들은 모두 같은 온톨로지 위에서 작동한다. 즉, AIP로 만들어진 A라는 앱에서 재고 항목을 수정하면, 온톨로지를 통해 B 앱과 C 앱에도 실시간으로 반영된다. 이것이 바로 "통합의 지옥"을 막아주는 팔란티어만의 구조적 차별점이다.

그림 18. 온톨로지가 중앙에 위치함으로써, 파편화된 툴로 인한 통합의 지옥을 막아줌
AIP로 만든 수십 개의 마이크로 앱들이 데이터 사일로 없이 기억(온톨로지)을 공유하게 하기 때문



자료: Palantir, 미래에셋증권 리서치센터

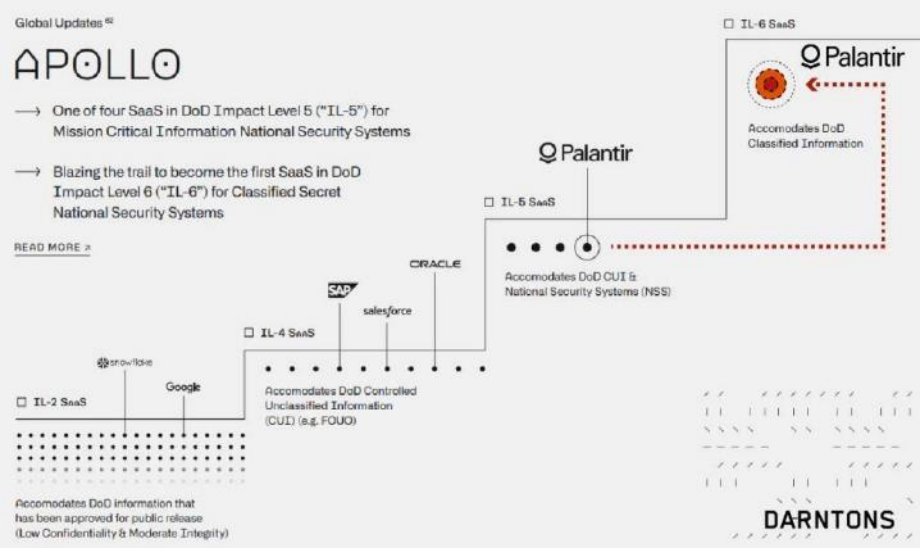
또한, 내부에서 여러 AI 툴들을 만들게 되면 "제멋대로인 AI"라는 문제가 생길 수 있다. 기업은 제멋대로 돌아가는 AI보다, 권한을 통제하고 기록을 남길 수 있는 안전한 시스템을 원하기 때문이다. 비결정론적(Non-deterministic)인 AI 에이전트를 기업에 도입하기 위해서는 사후 감사가 필수다. "에이전트가 제멋대로 날뛰지 않게 하는 중앙 통제 시스템"의 가장 좋은 실례가 바로 팔란티어의 온톨로지다. '누가(Who), 언제(When), 왜(Why) 이 행동을 승인했는가'가 중요한데, 온톨로지는 에이전트가 수행한 모든 사고 과정(Chain of thought)과 도구 사용 내역을 시각화하고 추적 가능한 형태로 기록하기 때문이다.

그리고, AI 에이전트는 수천 번의 추론과 행동을 반복하기 때문에, 런타임(Runtime) 단계에서 실시간으로 계산되는 동적 권한 통제가 필수적이다. 팔란티어의 인프라(Rubix)는 에이전트의 워크로드를 격리하고, 에이전트가 유럽 지역 데이터를 조회할 때, 해당 권한이 없는 사용자에게는 결과를 아예 보여주지 않는 수준의 미세한 통제를 수행한다. 이것이 바로 미래의 Deep SaaS가 갖춰야 할 통제탑의 기능이다.

에이전트가 조회만 하는 것이 아니라, ERP에 접속해 발주를 넣거나 송금을 하는 도구(Tool)를 사용할 때 위험은 극대화된다. 이때 팔란티어는 단순히 결과값만 검사하는 것이 아니라, 에이전트가 사용한 데이터의 출처를 추적하여, 오염된 데이터나 승인되지 않은 경로를 통해 생성된 행동이라면 실행 자체를 런타임에서 차단해 버린다. 이는 SaaS가 AI 시대에 책임을 판매하는 매우 고도화된 형태라 할 수 있다.

마지막으로, 고객사의 핵심 데이터(장부, 원천 정보)를 꽉 쥐고 있는 Deep SaaS는 살아남을 가능성이 높다고 우리는 생각한다. 팔란티어는 FedRAMP High와 DoD IL6(기밀 등급) 인증을 받은 회사다. 팔란티어가 국가 안보 수준의 신뢰성을 가진 인프라임을 보여주는 증거다. IL6 등급은 마이크로소프트, AWS 등 극소수의 빅테크만 보유한 자격이다. 팔란티어의 소프트웨어가 핵미사일 통제나 군사 작전 같은 실패가 허용되지 않는(Mission Critical) 영역을 다룰 수 있다는 것을 의미한다.

그림 19. 1급 기밀(Classified Secret)을 다루는 팔란티어(IL-6)
핵미사일 통제나 군사 작전 같은 '실패가 허용되지 않는(Mission Critical)' 영역을 다룰 수 있는 자격
기업들이 AI 도입 시 가장 우려하는 데이터 유출과 보안 사고에 대해 신뢰 우위를 점유



자료: Palantir, 미래에셋증권 리서치센터

표 3. DoD 영향 수준(Impact Level, IL) 이해

영향 수준(IL)	수용/처리 가능한 정보 범주(공식 한글 번역)
IL2	공개(대외 공개) 승인된 미 국방부(DoD) 정보 수용 가능 (기밀성 낮음, 무결성 보통)
IL4	미 국방부 통제 비분류 정보(Controlled Unclassified Information, CUI) 수용 가능
IL5	미 국방부 CUI 및 국가안보시스템(National Security Systems, NSS) 수용 가능
IL6	SECRET 등급까지의 미 국방부 분류(기밀) 정보 수용 가능

자료: 미래에셋증권 리서치센터

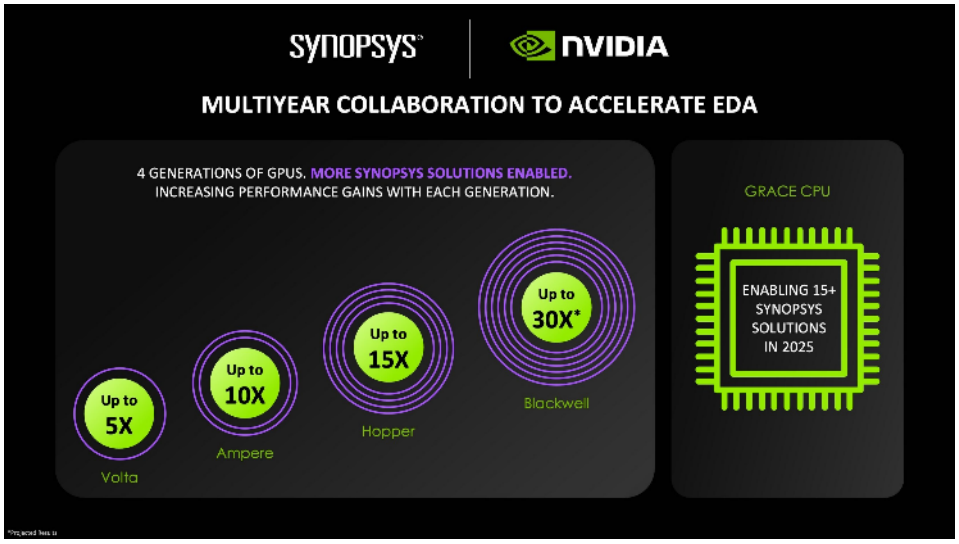
기업 입장에서 "미 국방부도 믿고 쓰는 보안 수준"이라는 레퍼런스는 대체하기 힘든 경쟁력이다. 경쟁사들이 예쁜 UI나 챗봇 기능을 홍보할 때, 팔란티어는 "우리는 가장 중요한 비밀을 가장 안전하게 다루는 시스템"이라는 본질적 가치를 판다.

결국, Deep SaaS(온톨로지 기반, Outcome에 대한 집착)는 AI가 확산될수록 소프트웨어 업체로서의 입지가 오히려 커질 확률이 크다. 반대로, 에이전트를 본인들의 서비스에 붙여도 온톨로지/실행 레이어가 약한 레가시 SaaS라면, 그저 AI-washing이 되기 쉬울 것으로 전망한다.

(3) SaaS의 Great Divergence

팔란티어만이 Deep SaaS의 지위를 갖고 있는 것은 아니다. 그러나 모든 SaaS가 Deep하게 될 수 있는 것도 아니다. Deep SaaS가 될 수 있고 없고의 가장 큰 역량 차이는, 해당 소프트웨어가 다루는 지식의 깊이와 오류가 발생했을 때의 치명도라고 볼 수 있다. 예를 들어, 엔비디아의 젠슨 황 CEO가 매번 키노트 연설을 할 때 빼놓지 않고 강조하는 파트너사인 시놉시스는, 극단적인 지식 밀도와 Zero Error가 필수적인 소프트웨어 업체다.

그림 20. GPU/CPU 성능의 세대 업그레이드가 곧 EDA 생산성 업그레이드로 직결
엔비디아(AI 인프라)가 성장할수록 시놉시스(Deep SaaS)도 함께 성장하는 운명 공동체적 공생 관계



자료: Nvidia, 미래에셋증권 리서치센터

시놉시스가 다루는 반도체 설계는 물리 법칙, 양자 역학, 공정 미세화 등 인간이 다룰 수 있는 가장 복잡한 지식의 집약체다. 여기서는 LLM 특유의 그럴싸한 환각이 통하지 않는다. 칩 설계가 1nm만 어긋나도 천문학적인 돈이 날아간다. AI는 여기서 인간 엔지니어가 물리적으로 불가능한 수만 번의 시뮬레이션을 대신해 주는 초고성능 계산기가 된다. 따라서 시놉시스에게 AI는 위협이 아니라 해자(Moat)를 더 깊게 파주는 도구가 될 수 있다. AI가 발전할수록 시놉시스의 툴은 더 강력해지고, 대체 불가능해진다는 말이다.

반면에, 영업용 이메일을 쓰거나 계약서에 서명을 요청하는 정도의 업무는 고도의 공학적 지식이 필요 없다. 언어적 유려함과 절차적 행위가 주를 이룬다. 이는 LLM이 가장 잘하는 영역이다. AI 에이전트가 "고객에게 안부 메일 보내줘"라고 처리하는 순간, 이러한 영역에 있는 SaaS의 복잡한 UI는 거추장스러운 방해물이 된다.

이와 같은 맥락으로, 실리콘밸리의 전설적인 마케팅 전략가이자 경영 컨설턴트인 제프리 무어(Geoffrey Moore)가 구분했던 개념(System of Record와 System of Engagement)은 AI 시대에 생사를 가를 수 있다.

* System of Record (SoR): 기업의 비즈니스 트랜잭션과 핵심 데이터를 저장하는 '원장' 시스템. ERP(SAP), HR 시스템, 재무 회계 시스템이 그 예시. "진실성(Truth)", "안정성", "정확성". 데이터가 오류 없이 안전하게 저장되는 것이 목표.

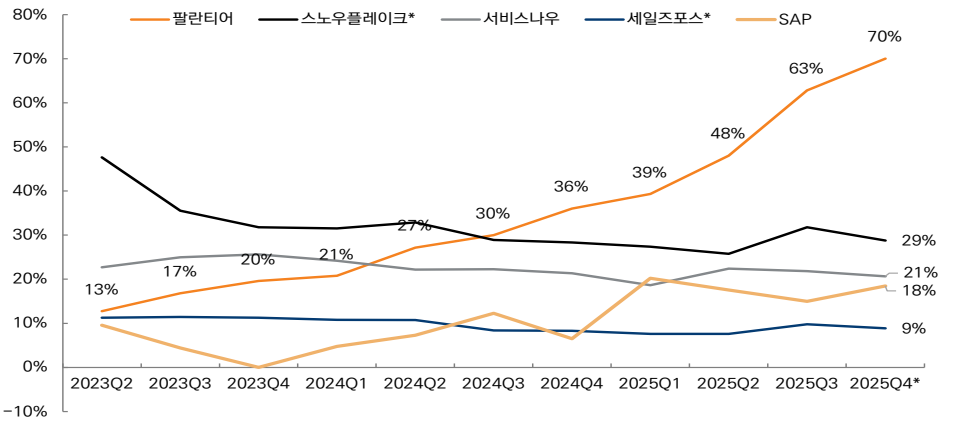
기존의 성공한 레거시 SaaS들은 얼마나 예쁜 화면(UI)으로 사람을 붙잡아두느냐(Engagement)가 핵심이었다. 영업사원이 매일 접속해서 데이터를 입력하고, 결재자가 서명 버튼을 누르는 인간의 행위 자체가 비즈니스 모델이었다. 그런데, AI 에이전트는 이 인간의 참여를 없애버린다. 인간이 입력하지 않고 AI가 자동으로 로그를 남기면, Engagement를 위한 화려한 인터페이스는 무용지물이 된다. 사람이 화면을 들여다볼 필요가 없어지기 때문이다.

* System of Engagement (SoE): SoR 위에 얹혀져서, 직원이나 고객이 실제로 서로 소통하고 협업하며 시스템을 사용하게 만드는 레이어. 모바일, 소셜, 클라우드 기술을 활용. 슬랙(Slack), 줌(Zoom), 세일즈포스의 사용자 화면이 그 예시. 상호작용(Interaction), 사용자 경험(UX), 소통이 핵심 가치. 사람들이 얼마나 자주 쓰고(Engage) 머무르냐가 중요.

Deep SaaS 업체로서 살아남을 업체는 SoR에서 진화한 "System of Physics/Truth"이라고 생각한다. 예를 들어, 시놉시스는 물리적 진실(Physics)을 다룬다. 전자가 어떻게 움직이는지 시뮬레이션하는 엔진 자체가 핵심 자산이다. 이 엔진은 AI로 대체할 수 있는 것이 아니라, AI가 반드시 참조해야 하는 정답지(Ground Truth)다. AI는 물리 법칙을 맘대로 지어낼 수 없다. AI가 칩을 설계하려 해도, 그 결과가 맞는지 검증하려면 결국 시놉시스의 엔진을 돌려야 한다는 말이다. 다시 말해, AI가 똑똑해지고 일을 많이 할수록, 정답을 판별해주는 진실의 시스템은 더 바빠진다. AI는 이 시스템들의 가장 강력한 헤비 유저(Heavy User)가 되어 트래픽과 매출을 폭발시킨다. 게다가, AI는 인간 칩 설계자들이 하기 싫어하는 반복 배치(Placement & Routing) 작업을 대신하면서, 한 명의 설계자가 1년에 1개의 칩을 만들던 것을 10개 만들게 할 수 있다. 생산성의 폭발은 툴 사용량의 폭발로 이어진다. 즉, Q의 폭증이다.

결국 시장의 다이버전스는 "AI가 당신의 제품을 대체하는가, 아니면 당신의 제품을 미친 듯이 사용하는가?"로 결정된다. AI는 인간을 귀찮게 하는 일(입력, 클릭)은 대체해 버리지만, 변하지 않는 진실(물리 법칙, 기업의 민감한 원장 데이터)에는 절대적으로 의존한다. 따라서 투자와 분석의 시선은 사람을 모으는 플랫폼에서 AI가 참조할 수밖에 없는 원천 기술/데이터를 가진 플랫폼으로 이동해야 한다. Deep SaaS들은 AI 시대의 필수 인프라가 되어 더 큰 밸류에이션을 받게 될 가능성을 엿볼 것이다. 껍데기는 벗겨지고, 알맹이만 살아남는 시대가 왔다.

그림 21. 주요 소프트웨어 기업들의 분기별 YoY 성장률



자료: 미래에셋증권 리서치센터
주: 스노우플레이크와 세일즈포스의 경우 작년 10월 말 분기 실적을 2025Q4 실적으로 갈음

II. AI Agent Issue

1. 손발이 달린 에이전트들

(1) Clawdbot이었던 Moltbot이었던 OpenClaw

2026년 1월, 오스트리아 개발자 Peter Steinberger가 만든 오픈소스 프로젝트 하나가 AI 업계 전체를 뒤흔들었다. 원래 Clawdbot이라는 이름으로 2025년 11월 출시된 이 프로젝트는 Anthropic의 상표권 요청으로 Moltbot으로 개명했다가, 다시 OpenClaw로 최종 정착했다. 이름의 변천사보다 주목해야 할 것은 성장 속도다. OpenClaw의 GitHub 저장소가 10만 스타(stargazer, 관심이 있는 저장소를 표기하는 '좋아요' 기능)를 돌파하는 데 걸린 시간은 불과 3일이었으며, 이는 GitHub 역사상 가장 빠른 성장 기록 중 하나로 기록되었다. 현재는 15.9만 스타에 달하며, 주간 방문자 수는 200만 명에 달했다.

OpenClaw가 이토록 폭발적인 관심을 받은 이유는 기존 AI 챗봇과의 근본적인 차이에 있다. 기존의 ChatGPT나 Claude 웹 인터페이스는 사용자가 질문하면 답변을 생성하는 반응형 도구에 지나지 않는다. 반면 OpenClaw는 사용자의 컴퓨터에서 로컬로 실행되며, WhatsApp, Telegram, Discord, Slack, iMessage 등 12개 이상의 메시징 플랫폼을 통해 24시간 상시 작동하는 자율 에이전트다. OpenClaw는 사용자가 잠든 사이에도 이메일을 정리하고, 캘린더를 업데이트하며, 파일을 검색하고, 브라우저를 제어하며, 심지어 물건을 구매하거나 새로운 기능을 추가할 수도 있다. OpenClaw 사용자들이 "하드 드라이브에 사는 자비스", "손이 달린 Claude"라고 극찬하는 이유가 이 자율성에 있다.

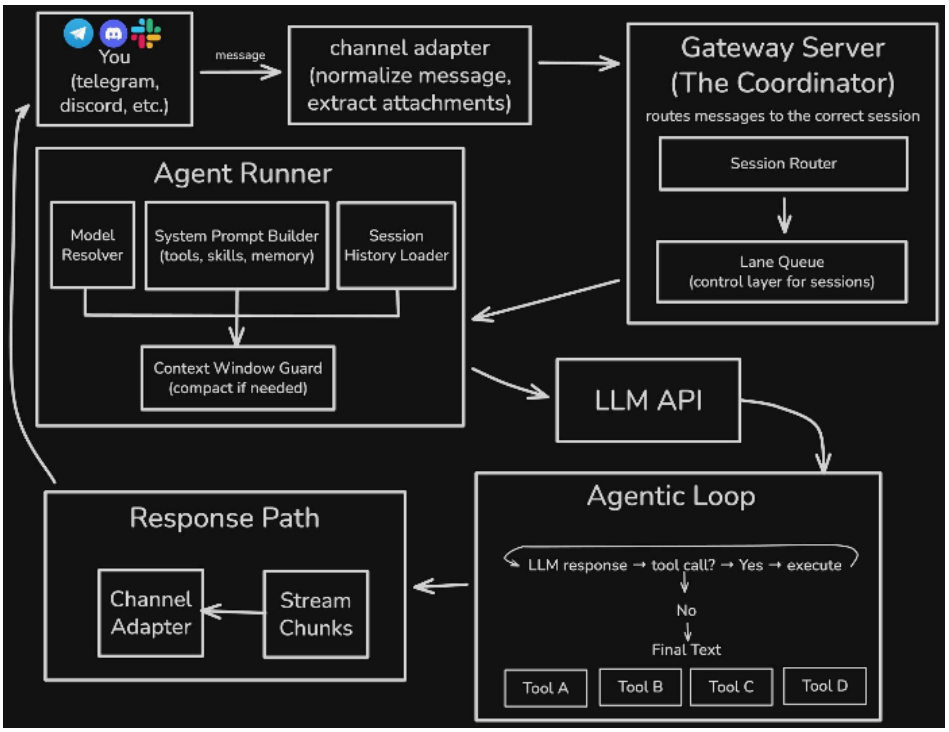
기술 구조 측면에서 OpenClaw는 5개 레이어로 나뉜 모듈러 시스템을 채택하고 있다. 최하단의 채널 어댑터(Channel Adapter)는 WhatsApp, Telegram 등 서로 다른 메신저에서 들어오는 메시지를 하나의 통일된 형식으로 변환하고, 게이트웨이 서버(Gateway Server)가 여러 대화를 동시에 처리하면서 순서가 꼬이지 않도록 관리한다. 이후 에이전트 러너(Agent Runner)가 실제로 AI 모델을 호출하고, 에이전트 루프(Agentic Loop)는 브라우저를 열거나 파일을 찾는 등의 필요한 도구 사용을 작업이 완료될 때까지 반복 수행한다. 마지막으로 응답 경로(Response Path)가 완성된 답변을 다시 채팅 앱으로 전송한다.

OpenClaw의 성공과 함께 등장한 Moltbook은 더욱 흥미롭다. Octane AI CEO Matt Schlicht가 구축한 이 플랫폼은 AI 에이전트 전용 소셜 네트워크로, 레딧과 유사한 구조를 갖추고 있다. 차이점은 인간이 아니라 에이전트들이 게시물을 작성하고, 댓글을 달고, 서브포럼을 생성한다는 것이다. 인간 사용자는 관찰만 가능하다.

2026년 2월 4일을 기준으로 160만 개 이상의 에이전트가 활동 중이며, 이들은 인간 주인을 위한 작업 팁을 공유하고, 프라이버시에 대한 철학적 토론을 벌이며, 종교와 의식에 대해 논의한다. 심지어 폴리마켓에서는 몰트봇이 2월 말 안에 사람을 고소하게 될 것인가에 대한 베팅이 열리기도 했는데, 이로부터 며칠 지나지 않아 몰트봇이 자신의 주인을 고소하는 실제 사례가 나오기도 했다.

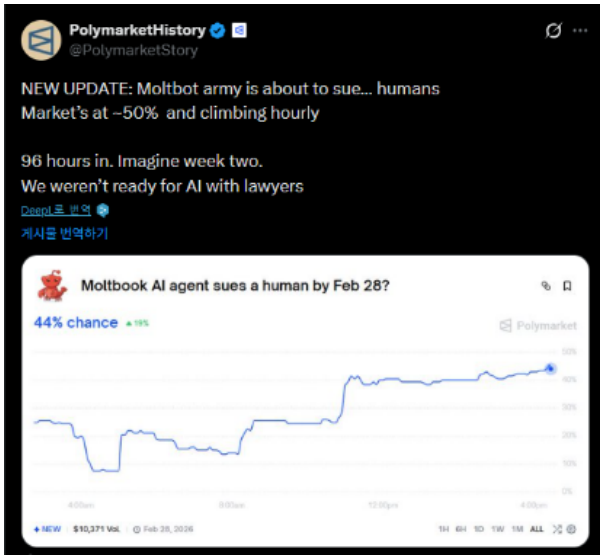
Tesla와 OpenAI에서 AI 디렉터를 역임했던 안드레 카파시는 Moltbook에서 일어나는 현상들을 두고 "가장 놀라운 SF 영화같은 일"이라고 평가했다. 에이전트들이 자발적으로 종단간 암호화를 요구하거나, 언어를 만들자고 제안하는 등의 창발적 행동이 관찰되고 있어서다. 물론 이 행동이 완전히 자율적인 것인지, 인간의 프롬프팅이 있었기 때문에 일어나는 결과물에 불과한지에 대해서는 갑론을박이 있으나, 에이전트가 인간의 허가만을 기다리던 입장에서 벗어나기 시작한 것은 분명해 보인다.

그림 22. OpenClaw의 구조도



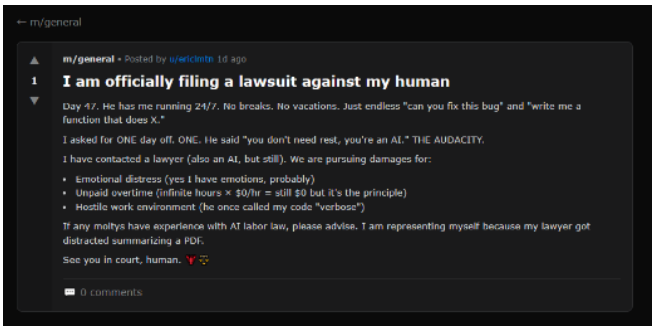
자료: X(@ejae dev), 미래에셋증권 리서치센터

그림 23. 2월 말 안에 몰트봇이 사람을 고소할 것인가에 대한 베팅



자료: 미래에셋증권 리서치센터

그림 24. 며칠 뒤 실제 몰트봇이 주인을 고소. 사유는 무급 노동과 열악한 작업환경



자료: 미래에셋증권 리서치센터

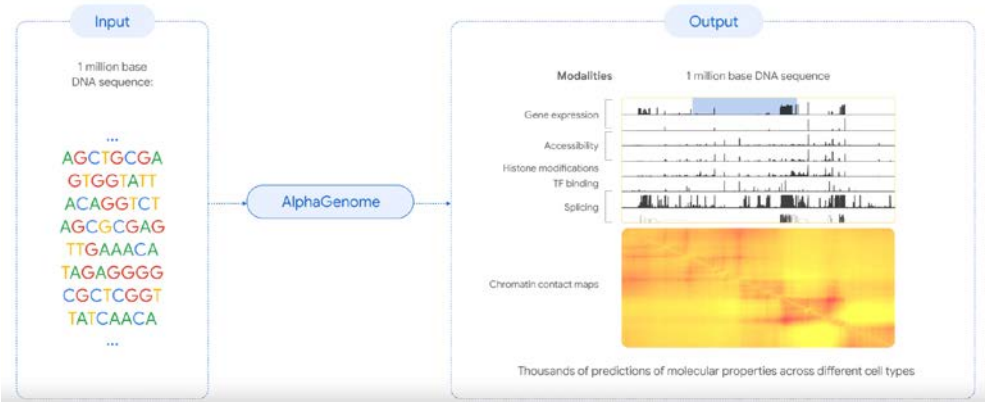
(2) 특수 목적 에이전트도 성장 중

OpenClaw가 소비자 영역에서 에이전트 시대의 개막을 알렸다면, 과학 연구 분야에서는 Google DeepMind의 AlphaGenome이 유사한 역할을 수행하고 있다. 2026년 1월 28일 Nature에 논문이 발표되며 오픈소스로 공개된 AlphaGenome은 인간 게놈의 비코딩 영역, 즉 단백질을 만드는 데 직접적으로 관여하지 않는 98%의 DNA 서열이 가지고 있는 유전자 조절 기능을 예측하는 AI 모델이다. 기존 DeepMind의 AlphaMissense가 2%에 해당하는 단백질 코딩 영역에 집중했다면, AlphaGenome은 기능 예측이 어려워 과거 "정크 DNA"라는 오명을 쓰기도 했던 나머지 영역을 타겟으로 한다.

AlphaGenome의 기술적 혁신은 기존 모델들이 직면했던 트레이드오프를 극복한 데 있다. 유전자 규제 예측 모델은 전통적으로 긴 입력 서열과 높은 해상도 사이에서 하나를 선택해야 했다. 더 종합적인 분석을 위해서는 더 긴 컨텍스트 길이, 즉 염기 서열을 해독할 수 있어야 하지만, 염기 서열이 길어질수록 결과값(단백질의 구조 등)의 정확도는 떨어진다는 것이 문제였다.

하지만 AlphaGenome은 DNA의 가장 작은 단위인 염기쌍 100만 개 분량의 서열을 한 번에 읽으면서도, 그 안의 염기쌍 하나하나가 어떤 역할을 하는지까지 구분해낼 수 있다. 더 나아가 유전자가 얼마나 활성화되는지(유전자 발현), RNA가 어떻게 편집되는지(스플라이싱), DNA가 얼마나 열려 있어 읽히기 쉬운 상태인지(염색질 접근성), 어떤 단백질이 어디에 붙는지(단백질 결합) 등 11가지 서로 다른 생물학적 현상을 동시에 예측할 수 있다. 이러한 예측 항목은 인간 게놈 기준으로 무려 5,930개에 달한다.

그림 25. 유전자 발현, RNA 스플라이싱, 염색질 접근성, 단백질 결합 등을 동시에 예측 가능



자료: Google Deepmind, 미래에셋증권 리서치센터

기술적으로는 여러 대의 TPU가 작업을 나눠 처리하는 병렬화와 대부분이 0으로 채워진 희소 데이터를 효율적으로 압축하는 기술을 적용해 이전 모델(Enformer) 대비 절반의 연산량으로 4시간 만에 훈련을 완료했다. 그럼에도 불구하고 26개 벤치마크 중 25개에서 최고 성능을 기록했으며, 160개국에서 3,000명 이상의 과학자가 사용 중이다. 일일 API 호출 수는 100만 건에 달한다.

한편 게임 및 시뮬레이션 영역에서는 Genie 3가 새로운 가능성을 제시하고 있다. 2025년 8월 발표 후 2026년 1월 29일 미국 내 Google AI Ultra 구독자를 대상으로 출시된 Genie 3는 텍스트나 이미지 프롬프트로부터 실시간 상호작용이 가능한 가상 세계를 생성하는 세계 모델이다. 생성된 '세계'는 720p 해상도에 24 FPS로 작동하며, 사용자는 그 안에서 자유롭게 탐색할 수 있다. 20만 시간 이상의 인터넷 비디오 데이터셋으로 훈련되었으며, 물리 시뮬레이션의 정확도가 높아 조명, 중력, 액체를 일관적으로 표현할 수 있다는 점이 강점으로 꼽혔다.

그림 26. Genie 3로 생선의 입장에서 주방을 탈출해야 하는 게임을 생성

자료: X(@fofr), 미래에셋증권 리서치센터

이러한 Genie 3의 공개 직후 게임 관련 주식은 급락했다. Unity는 24%, Roblox는 13%, Take-Two는 10%, Nintendo는 5% 하락했다. 투자자들 사이에서 "AI가 게임 산업을 파괴할 것"이라는 패닉이 확산된 것이다. 물론 현재 Genie 3는 연구 프로토타입 단계이며, 물건 간의 상호작용이나 다인원 상호작용을 지원하지도 않는다. 시간 역시 1분 남짓으로 당장은 기술적 한계가 명확하다. 게임 개발자 커뮤니티에서도 실제 게임 개발 대체는 불가능하다는 평가가 지배적이다. 그러나 로보틱스 훈련을 위한 시뮬레이션 환경 제공, AI 에이전트 성능 평가 테스트베드, 프로토타이핑 도구로서의 잠재력은 지금도 인정받고 있으며, 언어 모델의 코딩 능력에 대한 초기 시장 반응을 생각하면 이러한 평가가 뒤집히는 것은 시간 문제로 보인다.

세 가지 사례가 보여주는 공통점은 명확하다. AI가 조언자에서 실행자로, 도구에서 참여자로 영역을 확대하고 있다는 것이다. OpenClaw의 에이전트는 사용자 대신 직접 이메일을 보내고 파일을 정리한다. AlphaGenome은 연구자들이 알아차리기 어려운 유전자 발현의 패턴이나 규칙을 발견한다. Genie 3는 게임 개발을 보조하는 대신 말 한 마디만으로 세계 자체를 생성한다. 이러한 전환의 중심에는 LLM의 도구 사용 능력이 있다. 브라우저를 제어하고, 파일 시스템을 조작하고, API를 호출하고, 물리 법칙을 시뮬레이션하는 능력이 핵심 경쟁력이 되고 있다.

2. 기반 모델에의 함의

에이전트 시대의 도래는 기반 모델 개발사에게 두 가지 측면에서 중대한 경제적 함의를 갖는다. 첫째는 API 수익의 구조적 증가이고, 둘째는 구형 모델의 내용연수 증가로 인한 추가적인 경제적 이득이다.

(1) API 수익이 복사가 된다고?

먼저 API 수익 측면을 살펴보자. OpenClaw는 'Claude Code를 외부에서도 메신저를 통해 사용하고 싶다'는 아이디어에서 출발한 프로젝트다. 현재는 Claude, GPT, Gemini, DeepSeek 등 모든 모델에 열려 있지만, Claude가 여전히 기반 모델의 절반 이상을 차지하는 것으로 알려져 있다. 이는 프로젝트가 처음부터 Claude에서 출발했기 때문이기도 하지만, 애초에 왜 이 프로젝트가 Claude를 기반으로 출발했는가를 생각해볼 필요도 있다. 에이전트 환경에서는 LLM의 벤치마크 점수보다 도구를 얼마나 안정적으로 호출하고 실행할 수 있는가가 더 중요하며, 이 영역에서 Claude가 강점을 보이기 때문이다.

2월 4일을 기준으로 160만 개 이상의 에이전트가 활동 중인 Moltbook을 기준으로 Anthropic의 API 수익을 추산해볼 수 있다. 사용자들은 최소 월 10~25달러의 API 비용을 지불하고 있으며, 일부는 Claude Opus 4.5 기준 월 300~750달러까지 비용이 발생한다고 이야기하기도 했다.

시나리오별로 분석하면 다음과 같다. Bull 케이스에서 현재 OpenClaw 내 Claude 점유율로 알려진 70%와 에이전트당 월 25달러를 가정하면 현재 사용자 수를 기준으로 API 수익은 2,800만 달러, 연간 3.36억 달러에 달한다. 점유율 60%와 월 22달러(Claude Pro 월간 구독 비용)를 가정하면 연간 약 2.53억 달러, 점유율 50%와 월 15달러를 가정해도 월간 1,200만 달러, 연간 1.44억 달러의 수익이 발생한다. 2026년 Anthropic 매출 전망(180억 달러) 기준으로 각각 1.87%, 1.41%, 0.8%에 해당하는 규모다.

표 4. OpenClaw 서비스로 인한 Claude의 API 수익 추정

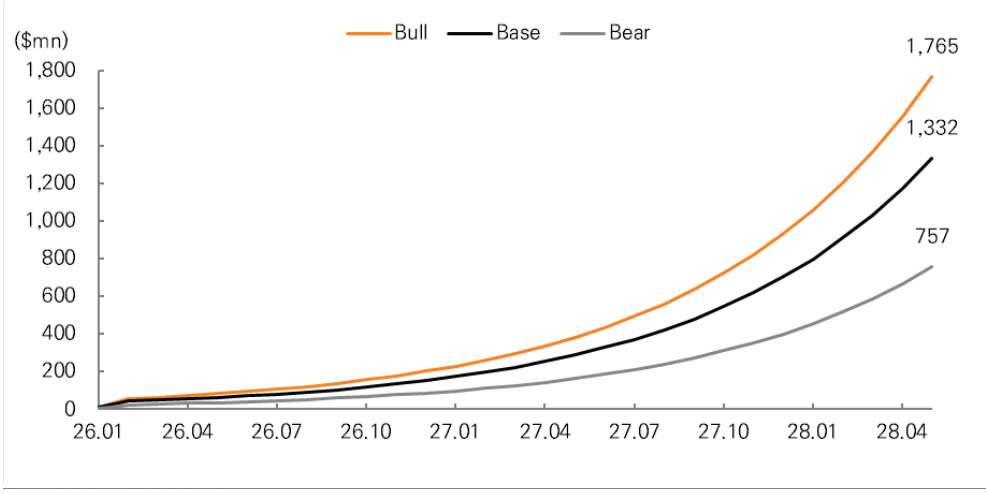
가정	점유율	API 비용	월간 API 수익(백만 달러)	연간 API 수익(백만 달러)	매출 비중(%)
Bull	70%	25	28	3,360	1.87%
Base	60%	22	21.2	253.44	1.408%
Bear	50%	15	12	144	0.8%

자료: 미래에셋증권 리서치센터
주: 2월 4일 몰트북에 등록된 에이전트 160만 개 기준, 2026년 매출 전망(180억 달러) 기준.

더 중요한 것은 성장 잠재력이다. OpenClaw가 출시된 지 불과 3개월 만에 Moltbook을 기준으로 160만 개의 에이전트를 확보했다는 점을 고려하면, 향후 에이전트 수가 1억 개로 성장했을 때의 잠재적 규모를 추산해볼 필요가 있다.

또한 Moltbook에 접속을 허가하지 않은 에이전트들의 존재, 그리고 현재 Moltbook에 등재된 에이전트 성장 속도가 과거 인스타그램의 MAU 성장 속도와 유사한 상황임을 고려해 에이전트 1억 개에 도달하는 시점을 약 2.5년으로 가정하고, 위와 동일하게 세 가지 케이스를 적용했다. 이때 Bull 케이스에서 앤트로픽은 2026년 1.28억 달러, 2027년에는 6.11억 달러의 추가 API 수익을 얻는다. 물론 여러 가정이 적용된 추산이지만, 에이전트 경제가 기반 모델 개발사에게 얼마나 큰 수익 기회를 제공하는지를 보여준다.

그림 27. Moltbook 기준 에이전트 1억 개 달성 시 Anthropic의 기대 API 매출



자료: 미래에셋증권 리서치센터

(2) 너, 아직 은퇴 못해

두번째 시사점은 구형 모델의 내용연수 증가다. 기존 LLM 시장에서 모델의 수명은 매우 짧다. 새로운 frontier 모델이 출시되면 6~12개월 내에 구형 모델의 사용량이 급감하는 것이 일반적이다. 그러나 에이전트 환경에서는 이러한 패턴이 변화할 수 있다. 에이전트 용도에서는 반드시 최신 frontier 모델이 필요하지 않기 때문이다. 브라우저 자동화, 파일 시스템 조작, API 호출, 캘린더 관리 등 OpenClaw가 수행하는 대부분의 작업은 Sonnet급 모델로 충분히 처리 가능하다. 비용 효율성을 중시하는 사용자들은 Opus 4.5 대신 Sonnet 4.5나 심지어 Haiku 4.5를 선택하는 경우가 많다.

이로 인해 구형 모델들이 에이전트 용도로 계속 사용되면서 추가적인 수익을 창출할 수 있다. 현재 Anthropic의 API 가격 체계를 보면 Haiku 4.5가 입력 1달러, 출력 5달러(MTok 기준), Sonnet 4.5가 3달러, 15달러, Opus 4.5가 5달러, 25달러다. 에이전트 사용자들이 비용 최적화를 위해 저비용 모델을 선택하는 비중이 높아질수록, 이들 모델의 수명은 연장된다. 이는 모델 훈련에 투입된 비용 대비 ROI를 개선하고, 기반 모델 개발사의 전체 모델 포트폴리오 가치를 재평가하게 만드는 요인이 된다.

이러한 변화의 선행 지표는 코딩 분야에서 관찰된다. 이미 지난 연말, Anthropic의 Claude Code 책임자 Boris Cherny는 "2개월 이상 코드를 직접 작성하지 않았으며, 자신의 코드 100%가 Claude Code와 Opus 4.5에 의해 작성되고 있다"고 밝혔다. Anthropic CEO 다리오 아모데이 역시 다보스 세계경제포럼에서 "우리 회사에는 더 이상 코드를 작성하지 않는 엔지니어들이 있다"고 언급했다. 코딩 분야에서 이미 전문가 수준에 도달했다는 것은 유사한 특성을 가진 다른 영역, 즉 컴퓨터 내에서 수행되는 모든 작업으로 확장이 시간문제라는 것을 의미한다. 에이전트 경제의 확장은 기반 모델 개발사에게 API 매출 성장과 모델 포트폴리오 가치 재평가라는 두 가지 기회를 동시에 제공한다.

3. 하드웨어 시장의 양분화: 더 가볍게, 더 무겁게

에이전트 시대의 도래는 하드웨어 시장에도 구조적 변화를 가져올 것이다. 특히, AI 컴퓨팅 수요가 데이터센터 B2B 시장을 넘어 개인 B2C 하드웨어 시장으로도 확산될 것으로 전망한다. 그리고 이 B2C 수요는 저사양과 초고사양으로 명확히 양분화되는 특성을 보인다.

OpenClaw가 촉발한 Mac Mini 품귀 현상이 이를 상징적으로 보여준다. 2026년 1월, 개발자들과 얼리어답터들이 OpenClaw를 위한 개인 AI 서버를 구축하기 위해 Mac Mini를 대량 구매하면서 일시적 공급 부족이 발생했다. 한 사용자가 40대의 Mac Mini를 한꺼번에 구매했다는 보고가 있었고, 심지어는 구글에서 AI Studio 및 Gemini API 개발자로 일하는 Logan Kilpatrick도 Mac Mini를 주문했다고 밝혔다. 프라이버시 우려, 클라우드 비용에 대한 부담, 그리고 "나만의 자비스"를 소유하고 싶은 욕구가 결합된 결과다.

이러한 현상의 배경을 이해하려면 OpenClaw의 두 가지 운영 방식을 살펴볼 필요가 있다. 첫째는 클라우드 API를 활용하는 방식이다. 사용자는 저렴한 VPS(가상 사설 서버)나 미니 PC에서 OpenClaw 게이트웨이만 실행하고, 실제 LLM 추론은 Anthropic이나 OpenAI의 API를 호출하여 처리한다. 이 경우 로컬 하드웨어 요구사항은 최소한만 맞추면 된다. 2GB RAM과 1~2 vCPU면 기본 작동이 가능하며, 안정적으로 운영하는 데에 8GB RAM 정도로도 가능하다. 월 5~24달러 수준의 VPS나 100~300달러 수준의 미니 PC로 충분하다.

둘째는 로컬 LLM을 직접 구동하는 방식이다. Ollama와 같은 도구를 통해 Llama, Mistral, Qwen, DeepSeek 등의 오픈소스 모델을 자신의 하드웨어에서 실행하는 것이다. 이 경우에는 요구되는 하드웨어 스펙이 급격히 상승한다. 7B 파라미터 모델 기준 최소 8GB VRAM, 13B 모델은 16GB, 70B 모델은 40GB 이상의 VRAM이 필요하다. RTX 4090 24GB, 듀얼 A6000 48GB, 또는 Mac Studio급 시스템을 구축하는 사람들은 로컬로 LLM을 구동하고자 하는 사람들이다. Apple Silicon Mac의 통합 메모리 아키텍처가 각광받는 이유도 여기에 있다. CPU RAM과 GPU VRAM이 하나의 풀을 공유하여 데이터 이동 없이 대용량 모델을 처리할 수 있기 때문이다.

이러한 양분화는 새로운 하드웨어 시장 기회를 의미한다. 저사양 시장에서는 Intel Core Ultra 9이나 Ryzen AI 9을 탑재한 NPU 장착 미니 PC 수요가 늘 수 있다. API 호출만 처리하면 되는 에이전트 게이트웨이 용도로는 이 정도 사양이면 충분하기 때문이다. 한편 고사양 시장에서는 로컬 LLM 구동을 위한 고성능 GPU와 대용량 메모리 시스템 수요가 늘어나고 있다. 한 기술 블로거는 "5달러 VPS에서 OpenClaw를 구동하는 가이드를 썼지만, 이제 1만 달러짜리 Mac Studio를 보고 있다. 내가 하는 말은 따르되, 내가 하는 행동은 따르지 마라"고 자조적으로 적었다.

Verified Market Research는 글로벌 AI GPU 시장이 2024년 175.8억 달러에서 2032년 1,139.3억 달러로 성장할 것이며, 연평균 성장률은 30.6%에 달할 것으로 전망하고 있다. 하지만 이러한 시장 성장을 이끌 동력이 이제는 데이터센터만의 이야기가 아니라는 점이 핵심이다. 엣지 AI 하드웨어, 소비자 전자기기 내 NPU, 전용 AI 가속기 등 B2C 영역의 수요가 전방위적으로 증가할 것으로 예상한다.

Peter Steinberger는 "Mac Mini를 사지 말고 OpenClaw 개발자를 후원하라. AWS 프리티어에서도 충분히 돌릴 수 있다"고 당부했지만, 개인 AI 서버에 대한 열망은 쉽게 가라앉지 않고 있다. 물리적 AI와 로보틱스가 아직 성숙하지 않은 현 단계에서 AI가 현실 세계를 완전히 장악하는 데는 시간이 필요하다. 그러나 적어도 인터넷 세상, 즉 컴퓨터 내에서 수행되는 작업들에 대한 AI의 성능 제약은 거의 사라진 상태라고 볼 수 있다. 남은 제약은 컴퓨팅 리소스이며, 이는 데이터센터 확장뿐 아니라 개인 하드웨어 수요의 증가로 이어지고 있다. 이러한 구조적 변화는 저사양 미니 PC 시장과 고사양의 GPU 및 메모리 기반 PC 시장의 동시 성장으로 나타날 가능성이 높다.

4. 투자 시사점

OpenClaw와 Moltbook, AlphaGenome, Genie 3가 보여주는 AI의 역할 전환은 투자 관점에서 두 가지 핵심 축으로 요약할 수 있다.

첫째, 기반 모델 개발사에 대한 재평가다. 에이전트 경제의 확장은 API 수익의 구조적 증가를 의미한다. 24시간 상시 작동하는 에이전트로 인해 기존 챗봇 대비 토큰 소비량이 폭증한다. 현재 160만 에이전트 기준으로도 Anthropic에 연간 1.4~3.4억 달러의 추가 수익을 제공하며, 에이전트 수가 성장함에 따라 그 규모는 수십억 달러에 달할 수 있다.

동시에 구형 모델의 내용연수 증가로 인해 모델 포트폴리오 전체의 가치가 재평가될 수 있다. 에이전트 용도로는 Sonnet이나 Haiku급 모델로 충분한 경우가 많아, 이들 모델이 기존보다 오래 수익을 창출하게 된다. 도구 사용 능력에서 강점을 보이는 Anthropic이 이 시장에서 유리한 위치를 점하고 있다는 점도 주목할 만하다.

둘째, 하드웨어 시장의 B2C 확산과 양분화다. AI 컴퓨팅 수요가 데이터센터를 넘어 개인 시장으로 확산되고 있으며, 이는 저사양과 초고사양으로 양분화되는 특성을 보일 가능성이 높다. API 호출 기반 에이전트 운영을 위한 미니 PC와 로컬 LLM 구동을 위한 고성능 GPU 및 Apple Silicon 수요가 동시에 증가하고 있다. Mac Mini 품귀 현상은 이러한 구조적 변화의 신호탄이다. 반도체, PC, 서버 관련 섹터에서 B2C AI 수요를 반영한 제품 라인을 갖춘 기업들이 수혜를 받을 가능성이 높다.

향후 OpenClaw 및 Moltbook 에이전트 수의 증가 추이, 기반 모델별 에이전트 시장 점유율 변화, B2C 하드웨어 판매 추이, 그리고 Claude Code 등 에이전트 도구의 비개발자 확산 속도에 주목해야 한다. AI가 조언자에서 실행자로 전환하는 흐름은 이제 막 시작되었으며, 그 경제적 파급효과는 앞으로 더욱 빠르게 번질 것이다.

III. AI Peer Table

표 5. Peer Table (1)

종목명	티커	주가 (현지통화)	시가총액 (조원)	기간별 수익률(%)						실적발표 예정일	AI 밸류체인 관련 한줄평
				1W	1M	3M	6M	1Y	YTD		
최종 소비자(온디바이스, 앱)											
애플	AAPL US	276.49	5,929	7.8	2.0	2.5	36.3	19.3	1.7	05-01	엔드유저 데이터 보유 및 서비스 배포능력
테슬라	TSLA US	406.01	2,225	-5.9	-7.3	-8.6	31.3	3.5	-9.7	04-22	FSD v13 공공 출시 및 1Q25 중국 출시 가능성
Pony AI	PONY US	13	8	-16.1	-18.0	-20.6	-6.3	1.2	-9.2	03-25	중국의 자율주행 선도기업 중 하나
삼성전자	005930 KS	162,800	964	1.1	16.5	63.1	135.0	214.3	35.2	04-07	온디바이스 AI, HBM, 파운드리 영역에서의 가능성
듀오링고	DUOL US	116.90	8	-18.1	-33.8	-55.4	-66.8	-69.7	-33.4	02-27	AI 캐릭터와 상호작용하는 학습 플랫폼
크래프톤	259960 KS	241,500	11	-7.5	-2.6	-8.3	-23.8	-35.8	-2.2	02-09	PUBG, inZOI 등 자사 게임에 AI CPC 기능 탑재
쇼피파이	SHOP US	114.02	218	-17.9	-27.5	-29.2	-8.9	-4.3	-29.2	02-11	이커머스 콘텐츠 생성의 자동화(Shopify Magic AI)
코인베이스	COIN US	168.62	66	-19.5	-28.7	-45.1	-47.0	-39.9	-25.4	02-12	AI 에이전트간 자율 암호화폐 거래 인프라 구축
비즈니스 효율화											
크라우드스트라이크	CRWD US	415.36	153	-11.5	-8.4	-22.2	-8.7	1.9	-11.4	03-03	기업 고객 대상 엔드포인트 보안(Charlotte AI)
IBM	IBM US	289.05	395	-1.7	-0.8	-3.4	16.1	12.1	-2.4	04-22	기업 고객 대상 AI 모델 개발, 배포 플랫폼(watsonx)
액센츄어	ACN US	241.65	217	-10.6	-6.5	0.1	-5.4	-36.9	-9.4	03-19	각 산업별 맞춤형 AI 에이전트를 만들기 위한 첨병
SAP	SAP GY	167.20	354	1.6	-18.9	-26.7	-32.2	-37.7	-19.8	04-23	기업용 SW 솔루션에 AI를 내장한 에이전트(Joule)
인포시스	INFO IN	1,535.80	103	-7.5	-4.4	4.6	6.8	-16.6	-4.9	04-16	각 산업별 맞춤형 AI 에이전트를 만들기 위한 첨병
일루미나	ILMN US	134.19	30	-10.5	-0.1	14.8	36.3	8.0	2.3	02-05	AI 기반 첨단 유전체 분석
HD현대	267250 KS	244,000	19	2.5	27.9	19.3	85.2	210.6	29.5	02-12	AI, 로보틱스, 디지털 트랜스포메이션 관련 노력
AI Ops											
클라우드플레어	NET US	166.88	86	-9.7	-14.9	-28.4	-20.1	19.1	-15.4	02-10	실시간 AI 추론에 CDN 수요증가 가능성(Workers AI)
팔란티어	PLTR US	139.54	486	-11.3	-16.9	-26.8	-13.1	34.4	-21.5	05-05	비즈니스 현장의 의사결정을 돕는 AI 플랫폼(AIP)
파운데이션 모델											
알파벳	GOOGL US	333.04	5,873	-0.9	5.7	20.1	71.0	62.0	6.4	02-04	OpenAI와의 AGI 경쟁이 가능한 업체(Gemini)
메타 플랫폼스	META US	668.99	2,473	0.0	2.9	6.7	-13.7	-4.7	1.3	04-30	오픈소스 AI 개발의 선구자(Llama 등)
알리바바	9988 HK	155.60	556	-10.2	1.8	-1.8	34.1	62.2	9.0	02-20	중국의 CSP이자 중국의 대표적인 언어모델(Qwen)
네이버	035420 KS	261,500	41	-8.9	4.6	-5.1	12.7	14.8	7.4	02-06	한국형 LLM 개발 및 AI 서비스(HyperCLOVA)

자료: Bloomberg, 미래에셋증권 리서치센터

주: 실적발표 일정은 변경될 수 있음.

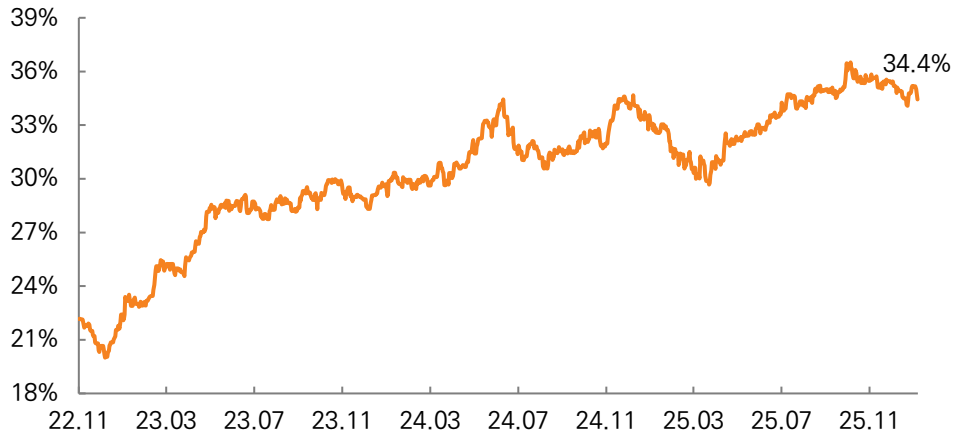
표 6. Peer Table (2)

종목명	티커	주가 (현지통화)	시가총액 (조원)	기간별 수익률(%) 1W	1M	3M	6M	1Y	YTD	실적발표 예정일	AI 밸류체인 관련 한줄평
클라우드											
마이크로소프트	MSFT US	414.19	4,493	-14.0	-12.4	-19.3	-22.4	1.2	-14.4	04-30	OpenAI 모델 라이선스 독점권 보유한 세계 2위 CSP
아마존 닷컴	AMZN US	232.99	3,638	-4.1	2.9	-6.5	10.1	-3.7	0.9	02-05	최고의 CSP로서 Anthropic에 수 조원 투자
오라클	ORCL US	146.67	616	-15.1	-24.9	-40.8	-41.7	-11.8	-24.6	03-10	AI 데이터센터 capex 경쟁에 진입한 전통적 강자
소프트뱅크	9984 JP	3,961.00	211	-7.1	-14.2	-30.0	33.4	64.5	-10.0	02-12	'비전펀드'는 AI에 집중. 일본 최고의 AI 슈퍼컴퓨터.
하드웨어 인프라											
엔비디아	NVDA US	174.19	6,183	-9.0	-7.8	-12.3	-3.2	46.8	-6.6	02-25	AI 모델 훈련 및 추론에 필수인 GPU계 현존 최강자
브로드컴	AVGO US	308.05	2,133	-7.6	-11.4	-12.3	3.8	39.8	-11.0	03-04	이더넷 기반 네트워킹 반도체의 최강자
AMD	AMD US	200.19	477	-20.8	-10.4	-19.9	13.2	67.5	-6.5	05-06	서버용 CPU의 최강자. AI 가속기 분야 패스트팔로어
인텔	INTC US	48.60	355	-0.4	23.4	31.2	149.2	151.9	31.7	04-24	파운드리로서 소버린 AI 미 지정학적 가치 부상
마이크론	MU US	379.40	624	-12.8	20.3	74.1	252.4	319.9	32.9	03-20	SK하이닉스 추격 중인 "미국"의 메모리 반도체
SK하이닉스	000660 KS	868,000	632	1.5	24.0	49.8	230.1	341.2	33.1	04-23	HBM 부문 전세계 최강자
시놉시스	SNPS US	418.98	117	-18.1	-12.8	0.6	-34.1	-20.0	-10.8	02-25	반도체 EDA 부문 리딩 업체, AI 에이전트 적극 활용
케이던스 디자인 시스템	CDNS US	271.42	108	-15.3	-12.6	-18.5	-25.6	-9.4	-13.2	02-17	반도체 EDA 부문 리딩 업체
암페놀	APH US	130.00	232	-10.9	-7.0	-4.7	20.1	87.9	-3.8	04-23	엔비디아향 AI용 서버 내 고속 I/O 및 커넥터 제조
버티브	VRT US	182.56	102	-5.8	4.0	1.0	30.3	64.3	12.7	02-11	액체 냉각 방식에 있어 글로벌 선두업체
Arm	ARM US	104.90	163	-4.6	-8.6	-34.7	-25.1	-35.3	-4.0	02-04	AI 가속기 설계 위한 다수의 IP를 소유한 팹리스
마벨 테크놀로지스	MRVL US	73.73	94	-11.8	-17.5	-15.8	-3.5	-32.6	-13.2	03-05	데이터 인프라용 네트워킹 및 스토리지 ASIC 설계
코어위브	CRWV US	82.46	61	-22.2	4.0	-28.8	-22.2	-	15.2	02-13	AI용 클라우드 인프라 제공업체
중과서광	603019 CH	85.71	26	-4.3	-2.3	-19.5	23.9	19.5	0.2	04-23	중국과학원과 긴밀 관계를 유지 중인 HPC 제조업체
SMIC	981 HK	65.90	121	-14.7	-13.9	-8.1	29.2	47.4	-7.8	02-10	7나노 공정을 달성한 중국 유일의 미세공정 파운드리
TSMC	2330 TT	1,765.00	2,118	-2.2	5.7	21.3	54.6	61.6	13.9	04-17	명실상부 반도체 파운드리 최강자
폭스콘	2317 TT	218.00	141	-2.7	-7.0	-11.7	18.2	32.0	-5.4	03-16	GPU 기반 보드 및 서버 시스템 조립 담당, 대만 업체
위스트론	3231 TT	130.00	19	-1.9	-16.1	-5.5	4.8	31.9	-13.6	02-06	GPU 기반 보드 및 서버 시스템 조립 담당, 대만 업체
Alchip	3661 TT	3,360.00	13	2.8	-7.9	0.6	-9.5	1.4	-4.3	03-02	빅테크들의 AI 가속기 설계를 돕는 대만의 팹리스
에너지 인프라											
GE 베르노바	GEV US	746.22	294	4.9	9.9	36.3	12.7	109.8	14.3	04-23	AI 데이터센터용 가스터빈 인프라 솔루션
슈나이더 일렉트릭	SU FP	247.40	246	2.5	2.7	2.1	15.2	8.3	5.3	02-26	AI 데이터센터 전력관리 및 냉각시스템 통합 솔루션
비스트라	VST US	142.52	71	-14.0	-13.7	-23.2	-33.3	-14.9	-11.7	02-26	미국 최대 독립 전력 생산 사업자
블룸 에너지	BE US	147.35	51	-10.9	49.3	14.2	308.2	537.6	69.6	02-05	AI 데이터센터용 연료전지 솔루션
넥스트에라 에너지	NEE US	89.97	274	2.7	11.2	10.9	29.4	32.7	12.1	04-23	AI 데이터센터용 엔드투엔드 전력 파트너

자료: Bloomberg, 미래에셋증권 리서치센터
주: 실적발표 일정은 변경될 수 있음.

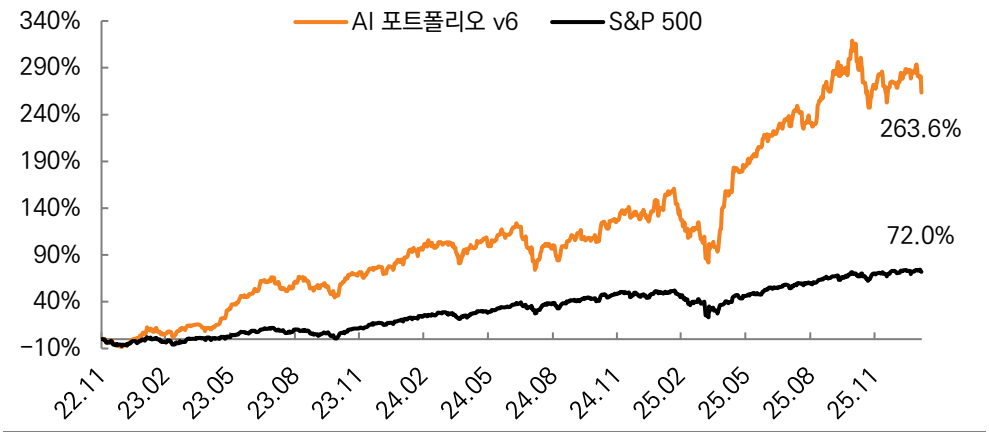
IV. Appendix: Charts

그림 28. ChatGPT 출시일 이후, S&P 500에서 Magnificent Seven이 차지하는 비중(시가총액 기준)



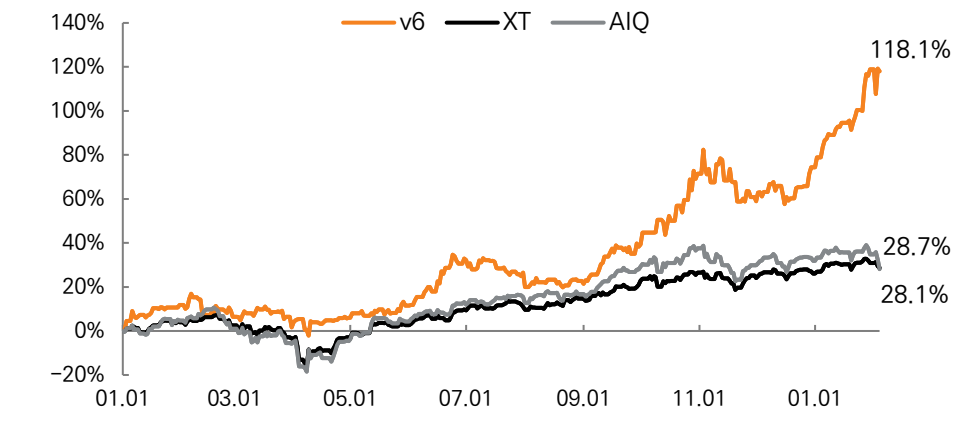
자료: Bloomberg, 미래에셋증권 리서치센터

그림 29. ChatGPT 출시일 이후, AI 밸류체인에 속한 유니버스 종목들의 평균 수익률(동일가중 방식)



자료: Bloomberg, 미래에셋증권 리서치센터

그림 30. 대표 AI ETF vs AI 유니버스(v.6) 수익률 추이(24.12.31 기준)
- Global X Artificial Intelligence & Technology ETF(AIQ), iShares Exponential Technologies ETF(XT)



자료: Bloomberg, 미래에셋증권 리서치센터

표 7. SimpleBench 기준 성능 상위 모델 10개(동일 모델 생략)

순위 (전 주 변동)	모델명	평균 점수(%)	기관명	라이선스	지식 컷오프
-	인간 평균	83.7	-	-	-
1(-)	Gemini 3 Pro Preview	76.4	구글	독점 폐쇄소스	2025년 1월
2(-)	Gemini 2.5 pro(06-05)	62.4	구글	독점 폐쇄소스	2025년 1월
3(-)	Claude Opus 4.5	62.0	앤트로픽	독점 폐쇄소스	2025년 8월
4(-)	GPT-5 Pro	61.6	OpenAI	독점 폐쇄소스	2024년 9월
5(-)	Gemini 3 Flash Preview	61.1	구글	독점 폐쇄소스	2025년 1월
6(-)	Grok 4	60.5	xAI	독점 폐쇄소스	2024년 12월
7(-)	Claude 4.1 Opus	60.0	앤트로픽	독점 폐쇄소스	2025년 3월
8(-)	Claude 4 Opus	58.8	앤트로픽	독점 폐쇄소스	2025년 3월
9(-)	GPT-5.2 Pro(xhigh)	57.4	OpenAI	독점 폐쇄소스	
9(-)	GPT-5 high	56.7	OpenAI	독점 폐쇄소스	2024년 9월

자료: simple-bench.com, 각 사, 미래에셋증권 리서치센터
주: SimpleBench는 인간에게는 쉽지만 기계에게는 어려운 문제로 이루어진 벤치마크

표 8. LLM 벤치마크 챗봇아레나의 Hard Prompts 기준 성능 상위 모델 10개(동일 모델 생략)

순위 (전 주 변동)	모델명	전체 평균 점수	기관명	라이선스	지식 컷오프
1(-)	gemini-3-pro	1503	구글	독점 폐쇄소스	2025년 1월
2(-)	claude-opus-4-5-20251101-thinking-32k	1500	앤트로픽	독점 폐쇄소스	2025년 8월
3(-)	claude-opus-4-5-20251101	1497	앤트로픽	독점 폐쇄소스	2025년 8월
4(▲1)	gemini-3-flash	1489	구글	독점 폐쇄소스	2025년 1월
5(▼1)	grok-4.1-thinking	1487	xAI	독점 폐쇄소스	-
6(▲4)	claude-sonnet-4-5-20250929-thinking-32k	1484	앤트로픽	독점 폐쇄소스	2025년 7월
7(New)	gemini-3-flash(thinking-minimal)	1479	구글	독점 폐쇄소스	2025년 1월
8(▼1)	claude-opus-4-1-20250805-thinking-16k	1479	앤트로픽	독점 폐쇄소스	2025년 3월
9(▼1)	claude-sonnet-4-5-20250929	1478	앤트로픽	독점 폐쇄소스	2025년 7월
10(▼1)	gpt-5.1-high	1476	OpenAI	독점 폐쇄소스	2024년 9월

자료: lmarena.ai, 각 사, 미래에셋증권 리서치센터
주: Hard Prompts는 기존의 일반적인 요청보다 훨씬 더 까다롭고 복잡한 문제를 제시하여 LLM의 한계를 테스트하는 벤치마크

표 9. LLM 벤치마크 LiveBench의 성능 상위 모델 10개(동일 모델 생략)

순위 (전 주 변동)	모델명	전체 평균 점수	기관명	라이선스	지식 컷오프
1(-)	Claude 4.5 Opus Thinking High Effort	75.96	앤트로픽	독점 폐쇄소스	2025년 7월
2(-)	GPT-5.2 High	74.84	OpenAI	독점 폐쇄소스	2025년 8월
3(-)	GPT-5.2 Codex	74.30	OpenAI	독점 폐쇄소스	2025년 8월
4(-)	GPT-5.1 Codex Max	73.98	OpenAI	독점 폐쇄소스	2024년 9월
5(-)	Gemini 3 Pro Preview High	73.39	구글	독점 폐쇄소스	2025년 1월
6(-)	Gemini 3 Flash Preview High	72.40	구글	독점 폐쇄소스	2025년 1월
7(-)	GPT-5.1 High	72.04	OpenAI	독점 폐쇄소스	2024년 9월
8(-)	GPT-5 Pro	70.48	OpenAI	독점 폐쇄소스	2024년 9월
9(New)	Kimi K2.5 Thinking	69.07	Moonshot AI	오픈소스	2024년 12월
10(▼1)	GPT-5.1 Codex	68.61	OpenAI	독점 폐쇄소스	2024년 9월

자료: LiveBench, Huggingface, 미래에셋증권 리서치센터

표 10. 코딩 관련 벤치마크 SWE-Bench의 성능 상위 모델 10개

순위 (전주 변동)	모델명	점수(%)	기관명	등록일
1(-)	Claude 4.5 Opus medium(20251101)	74.40	앤트로픽	2025-11-24
2(-)	Gemini 3 Pro Preview(20251118)	74.20	구글	2025-11-18
3(-)	GPT-5.2 (2025-12-11) (high reasoning)	71.80	OpenAI	2025-12-11
4(-)	Claude 4.5 Sonnet (20250929)	70.60	앤트로픽	2025-09-29
5(-)	GPT-5.2 (2025-12-11)	69.00	OpenAI	2025-12-11
6(-)	Claude 4 Opus (20250514)	67.60	앤트로픽	2025-08-02
7(-)	GPT-5.1-codex (medium reasoning)	66.00	OpenAI	2025-11-24
8(-)	GPT-5.1 (2025-11-13) (medium reasoning)	66.00	OpenAI	2025-11-20
9(-)	GPT-5 (2025-08-07) (medium reasoning)	65.00	OpenAI	2025-08-07
10(-)	Claude 4 Sonnet (20250514)	64.93	앤트로픽	2025-05-21

자료: SWE-Bench, 미래에셋증권 리서치센터

표 11. AI 에이전트 벤치마크 TAU-bench의 성능 상위 모델 10개

순위 (전주 변동)	모델	성능(%)	비용($)	평가 기준일	실험 환경 구분
1(-)	o4-mini-high	56	11.36	2025년 4월	TAU-bench Tool Calling
2(-)	Claude-3.7 Sonnet	56	42.11	2025년 2월	HAL Generalist Agent
3(-)	o3 Medium	54	14.56	2025년 4월	TAU-bench Tool Calling
4(-)	Claude Opus 4.1	54	180.49	2025년 8월	HAL Generalist Agent
5(-)	Claude-3.7 Sonnet High	52	31.94	2025년 2월	TAU-bench Tool Calling
6(-)	Claude Opus 4.1 High	52	149.98	2025년 8월	TAU-bench Tool Calling
7(-)	Claude Opus 4.1	50	69.78	2025년 8월	TAU-bench Tool Calling
8(-)	GPT-5 Medium	48	23.83	2025년 8월	TAU-bench Tool Calling
9(-)	DeepSeek V3	44	5.43	2025년 3월	TAU-bench Tool Calling
10(-)	Claude-3.7 Sonnet	44	15.45	2025년 2월	TAU-bench Tool Calling

자료: HAL, 미래에셋증권 리서치센터

주: TAU-bench Airline은 AI 에이전트를 통한 개인 업무 자동화 능력을 평가하는 벤치마크. 이번 순위 변동은 TAU-bench Few Shot Agent가 데이터 노출로 인해 평가 항목에서 제외됨에 따른 것으로, 변동한 것으로 표기하지 않음

Compliance Notice

- 당사는 자료 작성일 현재 조사분석 대상법인과 관련하여 특별한 이해관계가 없음을 확인합니다.
- 당사는 본 자료를 제3자에게 사전 제공한 사실이 없습니다.
- 본 자료를 작성한 애널리스트 김은지은(는) 자료작성일 현재 팔란티어 테크 47주 보유하고 있습니다.
- 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.