**Market Insight** | 2025.12.30

그림: Midjourney(레시피가 자동화되면 맛을 가려내는 안목이 권력이 된다)

# AI Bi-weekly

## 젠슨 황의 베팅은 Chain of Frame에서 나온다

**한종목**
chongmok.han@miraeasset.com

**김은지**
eunji.kim.a@miraeasset.com

CONTENTS

Highlight of the Week 3
기계적 숙련을 넘어 '취향'이 권력이 되는 시대 3
바이브 코딩 마찰 4

I. AI 시대에 살아남는 법: 실행의 종말과 안목의 시대 5
1. 2025년의 역설: 기술적 특이점과 대중의 공포 5
2. 인지적 재난: Opus 4.5가 촉발한 엔지니어링의 종말 7
3. 은폐된 진실과 노동 시장의 구조적 붕괴 8
4. 가치의 재구성: Wolf의 노동 3요소 이론과 실행의 종말 12
5. 인간 지성의 최후 성역: 젠슨 황과 'Chain of Frame' 15
6. 2026년을 위한 생존 전략: 안목(Taste)이 권력이다 18

II. 엔비디아의 Groq 자산인수 20
1. 구조적 병목과 규제 우회 20
2. Feynman 아키텍처: 이종 적층(Heterogeneous Integration) 22
3. 소프트웨어 해자: 결정론적 통제와 물리적 시간의 동기화 27
4. 메모리 계층의 재편과 경제성의 승리 30

III. AI & Coding 31
1. 확장하는 바이브 코딩 31
2. 변화하는 코딩 패러다임 32
3. 명과 암: 바이브 코딩에 대한 우려들 35
4. 매뉴얼이 없는 도구는 장인의 손을 탄다 39

IV. AI Peer Table 40

V. Appendix: Charts 42

# Highlight of the Week

## 기계적 숙련을 넘어 '취향'이 권력이 되는 시대

현재, 우리는 AI 역사상 가장 기이하고 역설적인 분기점에 서 있습니다. AI의 효용성은 증명이 되고도 남았지만, 정작 대중들이 AI에 느끼는 심리적인 지표는 스마트폰이나 인터넷 도입기보다 훨씬 낮은 수준으로 추락했기 때문입니다. 딥마인드의 데미스 하사비스가 지적했듯 바둑이나 예술처럼 인간 고유의 영역이라 믿었던 창의성이 데이터로 환원되면서 대중은 '신비의 상실'과 원초적 공포를 느끼고 있으며, 이는 AI를 단순한 혁신이 아닌 소셜 미디어 급의 '골칫덩어리'로 인식하게 만들었습니다.

이러한 공포는 기술 진보가 불러온 '실행(Execution)의 종말'에서 기인합니다. Anthropic의 모델 Opus 4.5와 코딩 에이전트 Claude Code의 결합은 기존 소프트웨어 엔지니어링에 대한 사실상의 사형 선고였습니다. 실리콘밸리의 엘리트 엔지니어들조차 자신이 더 이상 코드의 작성자가 아닌, AI가 쏟아내는 결과물의 검수자로 전락했음을 인정하고 있으며, 목표를 달성했다는 성취감보다는 자신의 존재 가치가 소멸했다는 '성취 후의 공허함'을 호소하고 있습니다. 이는 곧장 노동 시장의 구조적 붕괴로 이어질 가능성을 내포합니다. 이미 20대 초반 인력의 고용이 급감하는 '주니어 절벽' 현상이 현실화되었으며, 이는 인재 양성 파이프라인 자체가 끊기는 장기적인 사회 재난을 예고하고 있습니다.

이제 노동의 가치는 근본적으로 재구성되고 있습니다. Hugging Face의 창업자 토마스 울프가 분석한 대로, 도구를 다루고 과업을 수행하는 '실행' 능력의 시장 가치는 0에 수렴하고 있습니다. 누구나 AI를 통해 전문가 수준의 결과물을 즉시 만들어낼 수 있기 때문입니다. 이제 중요한 것은 AI가 파악하지 못하는 맥락을 읽어내는 '판단력'과, 불확실성 속에서 무엇을 만들어야 할지 결정하는 '주체성(Agency)', 즉 '안목(Taste)'입니다. 클라이언트의 모호한 요구를 해석하고 우리 회사의 결에 맞는 결과물을 골라내는 능력이 진짜 실력입니다.

인간 지성의 최후 성역은 어디에 있을까요? 젠슨 황의 직관, 즉 "나는 그냥 안다(I just know)"라는 발언에서 그 실마리를 찾을 수 있습니다. 이는 뇌과학적으로 설명 가능한 'Chain of Frame'입니다. AI는 텍스트나 이미지로 디지털화된 데이터는 완벽하게 학습하지만, 회의실의 미묘한 공기나 사람의 눈빛, 시대의 흐름 같은 비정형 데이터는 학습하지 못합니다. 인간의 뇌는 이러한 경험을 고차원적인 '장면'으로 압축해 저장하고, 결정적인 순간에 이를 비선형적으로 연결하여 최적의 해를 도출합니다. 젠슨 황이 데이터 상으로는 불가능했던 투자를 감행해 성공할 수 있었던 것은 바로 이 축적된 경험의 프레임 덕분입니다.

결론적으로 다가오는 2026년은 기계적인 실행의 시대가 가고, 인간의 고유한 안목이 권력이 되는 원년이 될 것입니다. 이제 개인에게 요구되는 인재상은 AI라는 무한한 실행 자원을 활용해 기획부터 결과물까지 홀로 완수하는 '1인 유니콘'입니다. AI에게 일을 맡기고 방관하는 것이 아니라, 자신의 안목으로 AI의 결과물을 끊임없이 평가하고 피드백을 주며 더 나은 결과를 도출해내는 '피드백 루프'의 설계자만이 살아남을 수 있습니다. 결국 기술이 발전할수록 역설적으로 가장 인간적인 경험과 취향, 그리고 안목이 가장 강력한 무기가 되는 시대가 도래한 것입니다.

## 바이브 코딩 마찰

2025년 12월, AI 코딩 도구 시장은 역사적 전환점을 맞이했습니다. Lovable은 5개월 만에 18억 달러에서 66억 달러로 3.67배 상승하며 3억 3천만 달러 투자를 유치했고, Cursor는 293억 달러 밸류에이션으로 자금 조달에 성공했습니다. GitHub Copilot은 2,000만 사용자를 확보했으며 Fortune 100 기업의 90%가 도입했습니다. Stack Overflow 설문조사에 따르면 개발자의 65%가 AI 코딩 도구를 주 1회 이상 사용하고, 41%는 작성하는 코드의 상당 부분이 AI에 의해 생성된다고 답했습니다. Polaris Market Research는 AI 코딩 어시스턴트 시장이 2024년 49억 달러에서 2032년 271억 달러로 성장할 것으로 전망했습니다.

하지만 이러한 폭발적 확장 이면에 균열이 드러나고 있습니다. OpenAI 창립 멤버 안드레이 카파시는 "규모 9 지진이 업계를 흔들고 있다"며 에이전트, MCP, 워크플로우 등 새로운 추상화 계층의 복잡성을 지적했습니다. 반면 Anthropic의 Claude Code 책임자 Boris Cherny는 30일간 259개의 PR을 AI로 완성하며 "코드는 더 이상 병목이 아니다"라고 선언했습니다.

현장의 반응은 엇갈립니다. Cursor CEO는 "바이브 코딩은 흔들리는 기반을 만든다"고 경고했고, 개발자 Simon Willison은 "증명 없이 코드만 던지는 것은 의무 태만"이라고 비판했습니다. Stack Overflow 설문에서 AI 도구에 대한 긍정적 감정은 70%에서 60%로 하락했으며, 72%는 "바이브 코딩은 내 방식이 아니다"라고 답했습니다. 불만의 1순위는 "거의 맞지만 완전히는 아닌" 솔루션으로, 66%가 이를 꼽았습니다.

고용 시장도 혼란스럽습니다. Stanford 연구에 따르면 22-25세 소프트웨어 개발자 고용은 2022년 대비 20% 감소한 반면, 26세 이상은 6-9% 증가했습니다. LA Times는 Stanford 졸업생들의 취업난을 보도했고, 한 졸업생은 "친구 절반이 5학년 석사를 하고 있다"고 증언했습니다. METR 연구는 더 충격적입니다. AI를 사용한 개발자는 실제로 19% 느려졌지만, 주관적으로는 20% 빨라졌다고 믿었습니다. Coinbase는 주니어의 코드 폭증이 시니어의 리뷰 부담으로 이어져 생산성 향상이 상쇄된다고 밝혔습니다.

게임 산업의 Clair Obscur 수상 취소 사건은 AI 사용의 투명성과 윤리 문제를 제기했습니다. 플레이스홀더로 사용된 AI 생성 텍스처가 논란이 되어 올해의 게임 수상이 취소됐고, 커뮤니티는 "AI는 도구일 뿐"과 "투명성 위반" 사이에서 분열했습니다. 숫자는 확장을 증명하지만, 현장은 투명성, 품질, 윤리의 경계를 둘러싼 균열을 목격하고 있습니다. 2025년은 프로그래밍 직업이 근본적으로 재정의되는 해이자, 동시에 그 변화의 방향성을 둘러싼 격렬한 논쟁이 시작된 해로 기록될 것입니다.

**표 1. AI 관련 주요 일정**

| 일 | 월 | 화 | 수 | 목 | 금 | 토 |
|---|---|---|---|---|---|---|
| 28 | 29 | 30 | 31 | 1 | 2 | 3 |
| | | | | | | |
| 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| | · CES 2026(~9) | | | · 삼성전자 실적(잠) | | |

자료: Bloomberg, 미래에셋증권 리서치센터

# I. AI 시대에 살아남는 법: 실행의 종말과 안목의 시대

## 1. 2025년의 역설: 기술적 특이점과 대중의 공포

### (1) 데이터로 증명된 '알파(Alpha)'와 대중 인식의 괴리

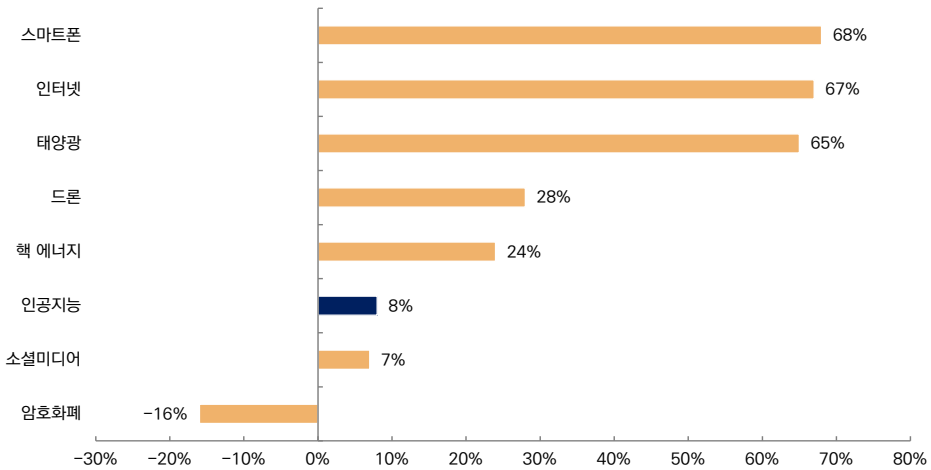
2025년 12월 30일 현재, 우리는 AI 역사상 가장 기이하고 이중적인 분기점에 서 있다. AI의 위력과 대중의 심리 사이에는 건널 수 없는 강이 흐르고 있기 때문이다.

우선, AI 효용성은 더 이상 논쟁의 대상은 아니다. Sparkline Capital의 창립자이자 CIO인 Kai W가 수행한 퀀트 분석 결과는 이를 입증한다. 섹터의 구분 없이, 심지어 보수적인 유틸리티나 금융 섹터 내에서조차 AI 인재를 채용하고 기술에 투자한 기업들이 그렇지 않은 기업 대비 유의미한 초과 수익(Alpha)을 달성하고 있음이 확인되었다. 경제적 유인의 실체가 숫자로 증명된 셈이다. AI가 제공하는 것이 100점이 아니라 '90점짜리 지식'에 불과하더라도, 다른 분야를 아예 모르는 비전문가들 입장에서는 'Zero to One'을 가능케 하는 강력한 레버리지가 되기 때문이다.

반면, 대중의 심리 지표는 처참하다. 2025년 여름, 미국인 2,300명을 대상으로 한 Searchlight Institute의 조사 결과는 AI에 대한 대중의 인식이 얼마나 위태로운지를 보여준다. AI가 사회에 미치는 영향이 '긍정적'이라고 답한 비율은 '부정적'이라고 답한 비율보다 고작 8%p 높았다.

이는 사회적 해악의 대명사인 '소셜 미디어'보다 단 1%p 높은 수치다. 인터넷(87% 긍정)이나 스마트폰(68% 긍정)과 비교하면 낙제점에 가깝다. 대중은 AI를 인정하면서도 소셜 미디어 급의 '필요악' 혹은 '골칫덩어리'로 인식한다는 것을 시사한다.

**그림 1. 주요 기술별 사회에 미치는 영향이 '긍정적(Net Positive)'이라고 답한 비율**




자료: Searchlight Institute, 미래에셋증권 리서치센터

### (2) 데미스 하사비스가 진단한 '신비의 상실'

이러한 대중적 반발의 기저에는 기하급수적 발전 속도에 대한 원초적 공포가 자리한다. 구글 딥마인드의 데미스 하사비스는 이를 두고 "신비의 상실"이라고 진단했다. 알파고가 바둑을 해석 가능한 데이터로 환원시키면서 수천 년간 이어진 바둑의 신비가 사라졌듯, 2025년의 AI는 글쓰기, 그림, 영상 등 인간 고유의 영역이라 믿었던 창의성을 기계적으로 복제해 내고 있다는 것이다. 이는 개발자들에게조차 경이로움과 상실감을 동시에 안겨주는 '달콤씁쓸한(bittersweet)' 딜레마를 안겨준다.

한편, 딥마인드의 AGI 연구원 Dileep George는 AGI(인공 일반 지능)와 AHI(인공 인간 지능), AUI(인공 범용 지능)를 명확히 구분하며, 우리의 목표는 인간의 생물학적 제약에서 자유로운 AGI라고 정의했다. 이는 얀 르쿤이 "인간은 물리 세계에 특화되었을 뿐 범용 지능은 없다"고 주장한 것에 대한 하사비스의 반박, 즉 "인간 지능은 전지전능하진 않지만 학습을 통해 무엇이든 익힐 수 있는 범용적 아키텍처"라는 논리와 맥을 같이 한다.

**표 2. 구글 딥마인드의 AGI 연구원 Dileep George가 제시한 세 가지 개념**

| 구분 | 정의 | 핵심 특징 | 실현 가능성 |
|---|---|---|---|
| AGI<br>(Artificial General Intelligence) | 인간과 유사한 수준의 일반 지능을 가진 인공 시스템 | • 인간 두뇌의 일반 원리를 기계로 구현<br>• 인간의 생물학적 제약에서 자유로움<br>• 무제한 메모리, 빠른 인터넷 접속 등 가능<br>• 인간보다 더 일반적일 수 있음 | 실현 가능(목표) |
| AHI<br>(Artificial Human Intelligence) | 인간 지능의 정확한 *복제품* | • 인간 두뇌의 생물학적 제약까지 포함<br>• 인간과 동일한 구현 방식<br>• 인간과 똑같은 한계를 가짐 | 굳이 만들 필요 없음 |
| AUI<br>(Artificial Universal Intelligence) | 무한한 능력을 가진 완벽한 지능 | • 무한한 컴퓨팅 파워 필요<br>• 무한한 데이터 필요<br>• 물리적으로 불가능한 조건 요구<br>• 영구 기관처럼 상상은 쉽지만 불가능 | 실현 불가능 |

자료: 구글, 미래에셋증권 리서치센터

이는 곧, 충분한 시간과 메모리(데이터)가 주어진다면, 인간이든 AI든 이론적으로 계산 가능한 모든 문제를 해결할 수 있다는 말이다. 따라서 AI는 이미 "극도로 범용적(Extremely General)"이기 때문에, AGI라는 것도 새로운 마법이 아니라 현재 모델의 연장선상에서 개발할 수 있다는 자신감으로 볼 수 있다.

## 2. 인지적 재난: Opus 4.5가 촉발한 엔지니어링의 종말

### (1) 실리콘밸리 엘리트들 "코딩은 죽었다"

2025년 12월, 소프트웨어 엔지니어링의 역사는 Anthropic이 내놓은 프론티어 AI 모델 Opus 4.5와 코딩 특화 서비스인 Claude Code의 결합 이전과 이후로 나뉜다고 봐야 할 것 같다. 이는 단순 업그레이드 수준이 아니기 때문이다. 전통적인 의미의 코딩이라는 행위에 대한 사형 선고이자, 엘리트 엔지니어들에게 닥친 인지적 재난에 가깝다.

**- Jackson Kernion (Anthropic의 파인튜닝 엔지니어):**
그는 "성취 후의 공허함"을 호소했다. 4년 전 입사 목표였던 'AGI 구축'이 Opus 4.5로 달성되었다고 느끼는 순간, 엔지니어링의 목적이 소멸했다는 것이다. 그가 "과학적 발견은 지능의 병목이 아니다"라고 덧붙인 것은 인간 존엄성을 지키기 위한 마지막 방어기제로 보일 정도다. 그는 이제 하드코어 엔지니어링이 아닌, AI에게 '성격'을 부여하고 심리를 분석하는 일종의 'AI 사회학'으로 눈을 돌리고 있다고 밝혔다. 이는 기술적 난제가 해결된 이후의 엔지니어들이 겪는 실존적 공허를 대변한다.

**- Boris Cherny (Anthropic의 클로드 코드 개발자):**
"30일 만에 4만 줄의 코드를 짰다. 모든 라인은 Claude Code가 썼다." 그의 발언은 압도적인 정복감을 드러내지만, 동시에 인간이 '작성자(Writer)'에서 '검수자(Reviewer)' 혹은 '방아쇠(Trigger)'로 전락했음을 시사한다. 그는 이를 "외계 기술(Alien and magical)"이라 표현했는데, 개발자조차 시스템의 내부 원리를 통제하지 못한다는 것을 자인하는 셈이다.

**- Andrej Karpathy(AI 분야의 전설적 엔지니어):**
AI 분야의 상징적 인물인 그조차 Claude Code에 비해서는 본인의 "Skill issue(실력 부족)"를 언급하면서, 심지어 "진도 9.0의 지진"이라고 표현했다. 결정론적 코딩의 세계가 확률론적 에이전트의 세계로 재편되면서, 20년간 쌓아온 지식이 무용지물이 되는 현기증을 느끼고 있는 것이다. 엔지니어링의 본질이 '구축'에서 블랙박스 AI를 달래는 '수습'과 '조율'로 변질된 것에 대한 깊은 우려다. 그의 속마음은 아마도 '이제는 프롬프트를 잘 던지고, 제멋대로인 에이전트를 달래는 게 '엔지니어링'이라고? 하지만 인정할 수밖에 없다. 결과물이 압도적이니까…' 정도로 이해할 수 있겠다.

**- Gianni Settino(AI 스타트업 창업자):**
그는 현재 상황을 "도구에 따른 계급 투쟁"으로 규정했다. Opus 4.5를 쓰는 자와 그렇지 못한 자의 격차는 단순한 속도의 차이가 아니라, '현실 인식'의 차이라는 것이다. "돈 아끼려고 싼 모델(Windsurf + GPT-4o) 쓰지 마라. Opus 4.5 안 쓰면 당신은 구석기시대 사람이다. AGI는 이미 여기 있다."라는 주장이다. 열등한 모델을 쓰면 단순히 작업 속도가 느린 게 아니라, 세상이 얼마나 변했는지조차 감지하지 못하게 된다는 경고다.

이들의 발언과 소회를 종합해보면, 겉으로는 기술의 진보를 찬양하지만 내면에서는 '순수한 프로그래밍'이라는 낭만의 시대의 종말을 직감하고 조용히 장례를 치르고 있는 듯 보인다. AGI 논쟁은 이미 끝났다. 기능적으로 AGI는 이미 여기 와 있다. 단지, 시대의 변화에 발맞춰 변화하지 않으면 아주 심각하게 뒤쳐진다(카파시가 말한 Fall behind)는 것은 확실해 보인다.

## 3. 은폐된 진실과 노동 시장의 구조적 붕괴

AI 기술의 발전이 인류에게 선사하는 장밋빛 미래, 그 이면에는 노동 시장의 붕괴가 진행되고 있다.

### (1) OpenAI 내부의 균열: 연구 윤리와 기업 선전의 충돌

2025년 하반기, OpenAI 내부에서는 파열음이 감지되었다. 1조 달러 규모의 IPO(기업공개)를 앞둔 시점에, 경제 연구팀(Economic Research Team) 소속 연구원들이 연쇄적으로 이탈하는 사태가 발생한 것이다. 이 사건의 중심에 선 톰 커닝햄(Tom Cunningham)의 폭로는 충격적이었다. 그는 사내 메신저를 통해 "경제 연구팀이 실제 연구를 수행하는 것에서 멀어지고, 대신 사측의 '선전 도구(propaganda arm)'처럼 행동하고 있다"고 강도 높게 비판했다.

**그림 2. OpenAI를 떠나 METR로 자리를 옮긴 톰 커닝햄의 X 계정**
**"경제 연구 팀이 실제 연구(real research)를 수행하는 것에서 멀어지고, 대신 사측의 '선전 도구(propaganda arm)'처럼 행동하고 있다"**

자료: 톰 커닝햄, 미래에셋증권 리서치센터

이 발언의 함의는 명확하다. 연구원들은 AI 모델이 노동 시장에 미칠 부정적 영향(일자리 대체, 소득 불평등 심화 등)을 가감 없이 발표하고자 했으나, 샘 알트만을 위시한 경영진은 이러한 '불편한 진실'이 대중에게 공개되는 것을 극도로 꺼렸다는 점이다. IPO를 앞두고 기업 가치에 치명적인 타격을 줄 수 있는 데이터들이 의도적으로 은폐되거나 축소되고 있다는 의혹이 제기된 것이다. IT 매체 Wired 역시 내부 소식통을 인용하여 OpenAI가 기술의 안전성과 긍정적 측면만을 강조하는 방향으로 연구를 통제하고 있다고 보도했다.

여기에 더해, 최고 커뮤니케이션 책임자(CCO)였던 Hannah Wong의 갑작스러운 퇴사는 이러한 의혹에 기름을 부었다. 통상적으로 IPO 직전은 스톡옵션 등 막대한 금전적 보상이 확정되는 시기다. 회사의 입장을 대변해야 할 CCO가 이 시점에 회사를 떠난다는 것은 단순 이직으로 보기 어렵다. 이는 내부적으로 윤리적 갈등이 있었거나, 혹은 다가올 미래에 대한 두려움이 작용해 침몰하는 배 혹은 폭주하는 기관차에서 뛰어내린 행위로 해석될 여지가 충분하다.

### (2) GPT-5.2와 GDPval 벤치마크: 인간 대체재의 공식화

OpenAI가 내부 잡음 속에서도 세상에 내놓은 GPT-5.2라는 명칭의 모델은 기존 모델들과는 차원이 다른 성능을 보여주며 노동 대체의 공포를 현실화하고 있다. 이를 증명하는 새로운 지표가 바로 'GDPval(Gross Domestic Product Evaluation)'이다.

GDPval은 단순한 코딩 테스트나 문장 요약 능력을 평가하지 않는다. 이는 3D 모델링 설계, 재무 모델링, 마케팅 계획 수립 등 인간 전문가가 수행하는 전체 프로젝트 단위의 업무를 AI가 처음부터 끝까지 수행할 수 있는지를 평가한다. 성능을 채점하는 전문 평가단은 해당 분야에서 최소 12년 이상의 경력을 가진 대기업 관리자급 인간 전문가들이다. 즉, 실제 채용과 승진을 결정하는 사람들의 눈높이에서 평가가 이루어진다.

GPT 모델의 성능 데이터는 극적인 반전을 보여준다.

- 2025년 9월: 당시 최고 모델이었던 GPT-5 High의 승률은 35%에 불과했다. 관리자들은 여전히 65%의 확률로 인간 직원의 결과물을 더 선호했다.

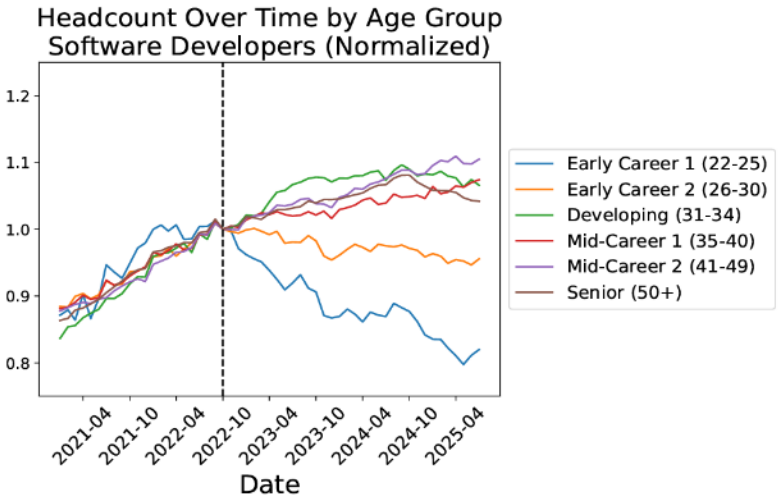
- 2025년 12월: GPT-5.2의 승률(무승부 포함)은 74.1%로 치솟았다. 순수 승률만 따져도 60%에 달한다.

이 데이터가 시사하는 바는 이제 기업 관리자 입장에서 12년 차 인간 전문가에게 일을 맡기는 것보다 GPT-5.2에게 일을 시키는 것이 결과물의 품질 면에서 더 낫다는 판단이 선다는 것이다. "전문 컨설팅 펌에서 팀 단위로 작업한 결과물 같다", "감사가 가능한(Auditable) 완벽한 스프레드시트"라는 평가단의 코멘트는, AI가 이제 보조 도구가 아니라 인간 노동의 대체재로서의 지위를 어느정도는 분명 획득했음을 의미한다.

### (3) 경제적 파장: '주니어 고용 절벽'과 화이트칼라의 위기

Anthropic이 스탠퍼드 대학과 협력하여 발표한 Anthropic Economic Index는 이러한 기술적 진보가 노동 시장에 미친 파괴적인 영향을 적나라하게 보여준다. 연구 결과는 Anthropic의 CEO 다리오 아모데이가 말했던 "화이트칼라 일자리의 피바다(Bloodbath)"라는 표현이 결코 과장이 아님을 증명한다.

**그림 3. 가장 심각한 고용 타격을 입은 계층은 갓 대학을 졸업하고 사회에 진입하려는 청년층**

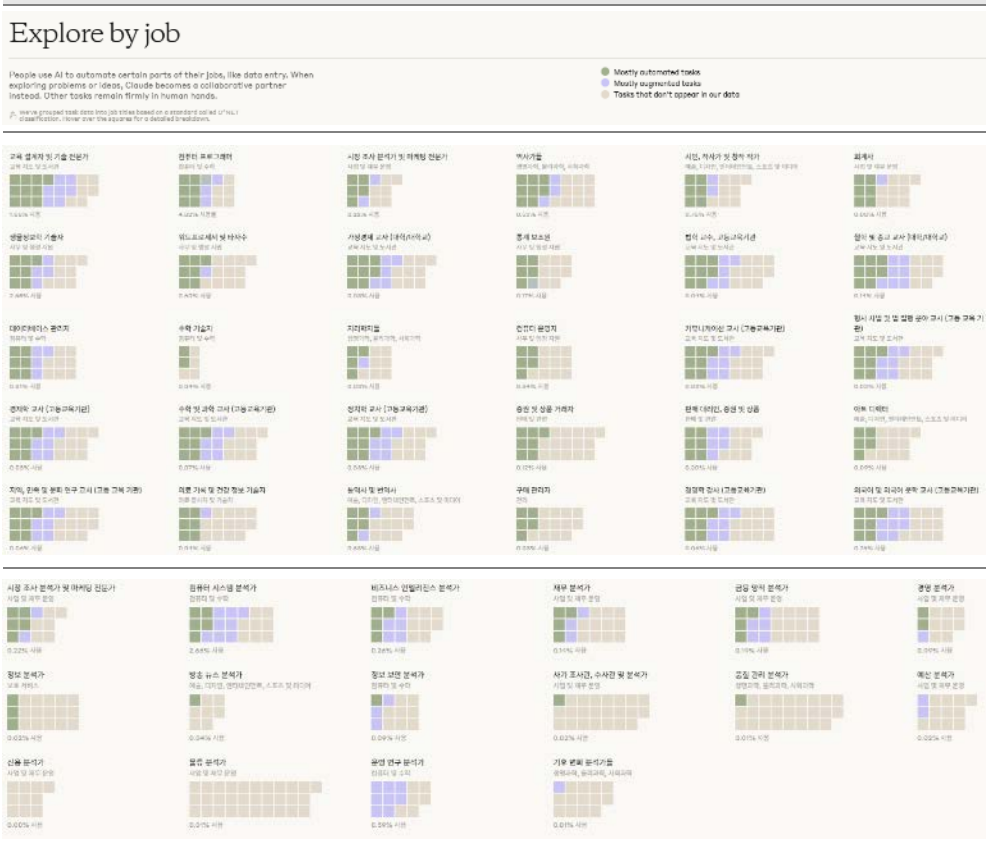



자료: Anthropic, 스탠포드 대학교, 미래에셋증권 리서치센터

Anthropic Economic Index를 보면, 소프트웨어 개발자, 데이터 웨어하우징 전문가, 웹 개발자, 카피라이터 등은 AI에게 대체될 가능성이 높은 '고위험군(자동화)'으로 분류되었다. 특히 심각한 것은 22세~25세의 초기 경력자(Early Career)들이 겪고 있는 주니어 절벽 현상이다. 2024년 말부터 2025년 중반까지의 데이터를 분석한 결과, 시니어(50세 이상)나 미드 커리어(30~40대)의 고용은 상대적으로 안정적인 반면, 초기 경력자의 고용 지수는 20% 이상 급락했다.

이유는 간단하다. 과거 기업들이 신입 사원에게 맡기던 이른바 기초 업(데이터 정리, 간단한 코딩, 초안 작성 등)를 이제 AI가 훨씬 더 빠르고 정확하며 저렴하게 수행하기 때문이다. 기업 입장에서는 비용을 들여 신입을 뽑아 가르칠 유인이 사라졌다. 이는 기술적으로 구현된 "사다리 걷어차기"다.

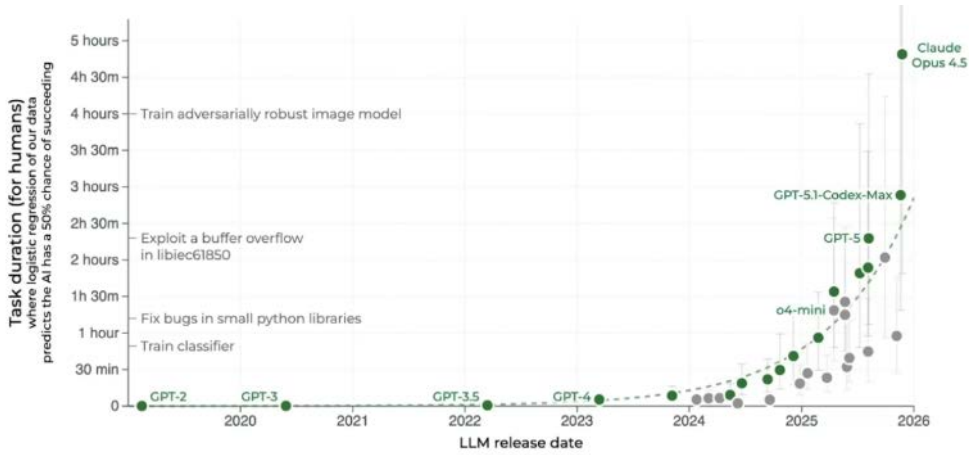
**그림 4. Anthropic이 말하는 직종별 자동화 위험도**
**초록색은 고위험군: 개발자, 데이터 웨어하우징 전문가, 웹 개발자, 카피라이터, 재무 분석가**
**보라색은 중위험군: 생산성을 극도로 높여 소수의 인원만으로도 업무가 가능해지는 영역**
**가장 아래의 이미지는 '애널리스트' 직군들로 필터링을 해본 위험도… 재무 분석가도 위험이 높다!**




자료: Anthropic, 미래에셋증권 리서치센터

또한, AI 발전 속도를 측정하는 METR 벤치마크에 따르면, Claude Opus 4.5는 인간이 약 5시간 걸리는 작업을 50% 확률로 성공해냈다. 심지어, 정치권에서는 이를 근거로 AI가 2027년까지 원격 근무 직종의 99%를 대체할 수 있다는 급진적인 시나리오를 담은 보고서를 내기도 했다.

**그림 5. AI 모델이 50% 성공률로 완수할 수 있는 작업"을 수행하는 데 "인간은 몇 시간이 걸리는가**
**즉, AI가 인간의 1시간짜리 일을 하는지, 5시간짜리 일을 하는지를 보는 것**
**2025년 말, Claude Opus 4.5는 인간이 약 5시간(4시간 49분) 걸리는 작업을 50% 확률로 성공**
**이를 기준으로 외삽을 해본다면, 2027년에는 엄청난 도약을 할 것**




자료: METR, 미래에셋증권 리서치센터

결론적으로, 우리는 지금 경기 침체라고 명명 짓기 보다는 노동의 구조적 종말(Post-Labor)을 목격하고 있다. 이미 자리를 잡은 경력자들은 AI를 도구로 활용하며 생존 시간을 벌었지만, 이제 막 사회에 진입하려는 청년 세대는 진입 장벽 자체가 AI로 대체되어 버린, 역사상 가장 불운한 세대가 될 위기에 처했다. 이들이 성장할 기회를 박탈당했다는 점은 향후 심각한 사회적 갈등과 불평등의 씨앗이 될 수 있다.

## 4. 가치의 재구성: Wolf의 노동 3요소 이론과 실행의 종말

Hugging Face의 공동 창립자이자 CSO인 Thomas Wolf는 현재의 노동 시장 혼란을 실행(Execution), 판단(Judgment), 주체성(Agency)이라는 세 가지 요소로 분해하여 설명한다. 그의 이론은 벤치마크 점수와 실물 경제 지표 사이의 괴리를 명쾌하게 해설한다.

### (1) The Paradox of 2025와 노동의 3요소

그런데, 2025년은 위에서 보다시피 기술적 특이점이 있지만, 일종의 실무적인 침묵이 공존하는 기묘한 해다. Claude Opus 4.5가 소프트웨어에 초점을 맞춘 가장 유명한 벤치마크인 SWE-bench에서 75%의 해결률을 기록했고, ChatGPT의 주간 활성 사용자는 거의 10억 명에 육박한 상황이다. 그러나 수많은 기업의 생성형 AI 프로젝트는 여전히 폐기되고 있다. 특히 Remote Labor Index라는 'AI의 원격 근무 테스트'에서 최고의 AI 에이전트 툴 중 하나로 꼽히는 ManusAI조차 성공률이 고작 2.5%에 불과하다는 사실은 충격적이다.

이러한 '벤치마크와 현실의 미스매치(Benchmark-Reality Gap)'는 우리가 일(Work)을 단순히 과업(task)의 나열로 오해한 데서 기인한다. Thomas Wolf는 이러한 미스매치를 설명하고 인간이 살릴 수 있는 기회를 설명하기 위해 노동의 가치를 세 가지 층위로 분해한다.

**- 실행(Execution)**

도구와 방법론을 마스터하고 주어진 과업을 수행하는 능력이다. 코딩, 번역, 요약 등이 여기에 해당한다. 이는 대개의 AI 벤치마크가 측정하는 영역이자, AI가 가장 완벽하게 정복한 영역이다. AI가 인간보다 더 빠르고 정확하게 수행하게 되면서, 실행 능력의 시장 가치는 0에 수렴하고 있다. 과거 인재의 척도였던 코딩 테스트 통과 능력은 이제 무의미해졌다.

**- 판단(Judgment):**

과업이 '왜' 중요한지, 회사의 거시적 목표와 문화, 가치관에 어떻게 부합하는지를 이해하는 능력이다. AI는 텍스트를 생성할 수 있지만, 그 텍스트가 회사의 미묘한 정치적 상황이나 브랜드 톤앤매너에 맞는지 느끼지 못한다. 판단력은 정적인 데이터가 아니라 동적인 맥락(Context)에서 나온다. "분위기 파악하고 선을 넘지 않는 능력"은 여전히 인간 관리자의 몫이다.

**- 주체성(Agency):**

불확실성 속에서 길을 찾는 능력으로 가장 중요한 능력이다. 다음에 무엇을 할지 예측하고, 제안하고, 때로는 "(명령을 받긴 했지만) 이 일을 하지 않는 것이 낫다"고 결정(Strategic Stop)하는 배짱이다. 이는 정해진 궤도 밖으로 나가는 상황에서 빛을 발한다. AI는 목적지를 주면 경로를 생성할 수 있지만, '지금 우리가 어디에 있고, 왜 저기로 가야 하는가?'라는 목적 함수 자체를 설정할 수는 없다.
Cursor의 시니어 디자이너인 Ryo Lu가 지적했듯, "코드를 짜는 건 몇 분이면 된다. 병목은 취향(Taste)이다." 모든 것이 자동화된 세상에서 유일하게 희소한 자원은 '무엇이 좋은 것인가'를 알아보는 안목이다. GPT-5 등 도구가 평준화되었다면, 이제 차별점은 도구를 쥐고 있는 사람의 미적/전략적 취향에서 나온다는 말이다.

**그림 6. 토마스 울프의 핵심 논지인 이상적 노동자는 세 가지 특성의 융합체임**
**'실행(Execution)'은 기술적 과업 수행 능력으로, 이제 AI에 의해 자동화되는 영역**
**'판단(Judgment)'은 회사의 목표와 가치라는 더 큰 맥락을 이해하는 지혜를 상징**
**'주체성(Agency)'은 불확실한 갈림길에서 방향을 설정하고 주도적으로 나아가는 결단력**
**AI가 '실행'의 가치를 급락시키면서, 인간의 경쟁력은 '판단'과 '주체성'의 교집합에 집중!**
**미래의 채용은 이 세 가지 영역을 모두 아우르는 인재, 즉 완성형 노동자를 찾는 전쟁이 될 것**

자료: Thomas Wolf, 미래에셋증권 리서치센터

### (2) 주체성(Agency)의 초격차가 낳는 것

Wolf의 이론을 실제 업무 시나리오에 대입해보면, AI가 대체할 수 있는 영역과 없는 영역의 경계가 명확해진다. 예를 들어, 소프트웨어 개발 팀 입장에서 생각해보자. 클라이언트가 "결제 버튼을 빨간색으로 바꿔달라"고 했을 때, AI(실행)는 군말 없이 코드를 수정하여 빨간색 버튼을 만든다.

그러나 기업의 맥락을 아는 인간 실무자라면, 클라이언트에게 "빨간색으로 바꾸면 사용자들이 '위험' 신호로 인식해 결제율이 떨어질 수 있다. 대신 우리 브랜드 컬러 중 강조색인 '주황색'을 쓰는 게 어떨까요?"라고 제안하며 수정할 수 있다.

나아가, 주체성(Agency)을 가진 고위급 관리자라면, "지금 결제율이 낮은 건 색깔 때문이 아니라 로딩 속도 때문이다. 디자인 수정이 아니라 결제 모듈 경량화를 해야 한다"며 클라이언트의 요청을 거절하고 새로운 판을 짠다. 이것이 대체 불가능한 인간의 가치다.

또 다른 예를 들기 위해, 서비스 장애가 생겼을 때 고객들에게 사과를 해야 하는 입장에 처해있다고 생각해보자. AI(실행)는 인간을 대신해 완벽한 문법의 사과문을 작성할 수 있다. 이는 매우 간단한 일이다.

분위기(뉘앙스)를 읽을 수 있는 실무자라면 "우리 주 고객층은 20대인데, 너무 딱딱하고 식상한 사과문을 올리면 오히려 비꼬는 댓글이 달릴 거야. 지금 경쟁사가 비슷한 이슈로 욕을 더 먹고 나락갔으니 좀 더 구체적으로 진정성을 담자."면서 문구를 작성할 것이다.

그 너머, 판단력(Judgment)을 가진 인간 리더라면 이런 주체성을 펼칠 수도 있다. "사과문만 올리지 말고요, 개발팀의 디버깅 현장을 생중계하여 우리의 진심을 보여주자"는 파격적인 기획을 제안할 수도 있는 것이다. 판을 뒤집는 직관(Taste)과 배짱은 인간의 것이다.

이처럼, 실행 가치의 추락과 주체성의 희소 가치 급등이라는 현상이 극대화될 것으로 사료된다. 하지만, 이러한 가치 이동은 노동 시장에 비대칭적 붕괴를 가져올 수밖에 없다. 위에서 살펴봤듯 Anthropic과 스탠포드 대학교의 분석 데이터에 따르면, AI 노출 직군에서 22~25세 노동자의 고용은 급감한 반면, 시니어급 고용은 오히려 증가하고 있다. 과거 신입 사원이 수행하던 실행 중심의 업무가 AI로 대체되면서, 주니어들이 조직의 맥락을 배우고 통찰력을 키울 기회 자체가 사라지는 것이다.

여기서 문제는, 시니어들이 갖고 있는 경험(Experience)이라는 것도 실패의 데이터베이스이자 맥락의 '라이브러리'라는 점이다. 현재의 효율성을 위해 신입 채용을 줄인 결과, 5~10년 뒤 판단력을 갖춘 시니어 리더를 공급할 인재 파이프라인이 끊어지는 사태가 올 수도 있다. 이는 기업과 사회 전체의 장기적 붕괴 시나리오가 아닐까?

결론적으로, 2030년으로 가면서 노동 시장의 변화는 미리 읽을 수 있다. 실행만 잘하는 인재는 이제 설 자리가 점점 없다. AI라는 무한한 실행 자원을 데리고 가치를 조립(및 Orchestrate)해낼 수 있는 판단력과 주체성, 그리고 압도적인 안목(Taste'을 가진 소수의 인재만이 살아남을 것이다. 실행은 상품(Commodity)이 되었고, 취향은 권력이 되었다.

## 5. 인간 지성의 최후 성역: 젠슨 황과 'Chain of Frame'

2025년의 AI 혁명 속에서, 엔비디아의 CEO 젠슨 황이 보여주는 행보는 인간 지성이 기계 앞에서 어떻게 살아남을 수 있는지, 아니 어떻게 기계를 압도할 수 있는지에 대한 결정적인 단서를 제공한다.

### (1) "I just know": 엑셀을 거부한 자의 특권

a16z가 진행한 인터뷰에서 SemiAnalysis의 편집장 딜런 파텔(Dylan Patel)은 젠슨 황의 경영 철학을 상징하는 한 문장을 전했다.

*젠슨 황: "나는 스프레드시트를 싫어한다. 안 본다. 그냥 안다*
*(I don't like spreadsheets. I don't look at them. I just know.)"*

이 발언은 천재의 허세는 아닐 것이다. Thomas Wolf가 제시한 노동의 3요소(실행, 판단, 주체성) 이론으로 완벽하게 설명되는, AI 시대 인간 리더십의 정수라고 우리는 생각한다.

**– 스프레드시트의 본질 = 실행(Execution):**

엑셀은 과거 데이터의 집합이자 '선형적 추론'의 도구다. 재고 회전율이나 ROI 분석과 같은 것들은 전형적인 실행 및 판단(Judgment)의 영역이다. 이 영역에서 AI는 엑셀을 인간보다 100만 배 더 잘 다룬다. 만약 젠슨 황이 실행이나 판단 레벨의 CEO였다면, 그는 엑셀 데이터를 신봉했을 것이다.

**– AI의 한계:**

AI에게 엑셀 데이터를 주고 분석을 시켰다면 결과값은 언제나 "투자 금지(Do not invest)"였을 것이다. 예를 들어, 엑셀에 의존했다면 4년 전, "암호화폐 시장이 폭락했으니 GPU 생산을 줄이십시오"라고 AI는 말했을 것이다. 하지만, 젠슨 황은 상황을 다르게 해석했다. "트랜스포머 모델이 AI의 가능성을 연다. 이건 컴퓨팅 패러다임의 전환이야. 지금 줄이면 AI 시대를 놓쳐. 재고고 뭐고 더 베팅한다."는 식이다. 물론, 가속 컴퓨팅이라는 시장은 2023년 11월 30일 이전에는 존재하지 않았고, 데이터가 없는 시장에 대한 투자는 엑셀상에서는 언제나 '비합리적'이었을 것이다.

**– 주체성(Agency)의 발현:**

이처럼, 젠슨 황이 엑셀을 거부한 것은, 혁신이 데이터의 연장선상에 있지 않음을 알았기 때문이다. 그는 AI가 대체할 수 있는 최적화'를 거부하고, AI가 절대 계산할 수 없는 비전의 베팅을 선택했다.

딜런 파텔이 젠슨 황의 베팅에 대해 언급한 "기능 검증 전 대량 생산(Full Wafer Order)" 전략은 주체성의 극치다. 젠슨 황은 "틀리면 수십억 달러를 날린다"는 엑셀의 경고를 무시하고, "경쟁사보다 6개월 먼저 시장을 장악하는 가치가 재고 상각 비용보다 크다"는 것을 주체적으로 결정했다.

결국 "I just know"는 마법이 아니라, 30년간 축적된 직관과 주체성, 그리고 안목(taste)의 결정체다. 그는 시장 조사를 하는 게 아니라, 자신의 Taste를 시장에 강요(Impose)했다. AI가 엑셀을 완벽하게 다루는 2025년, 엑셀을 덮고 본인의 통찰력으로 수조 원을 베팅하는 능력이야말로 인류가 기계 앞에서 가질 수 있는 최후의 성역이다.

### (2) Chain of Frame: 고차원 데이터의 생물학적 압축

그렇다면 인간이 가지는 직관 내지는 통찰이라는 것의 실체는 무엇인가? 뇌과학적 관점에서 이는 'Chain of Frame'이라는 개념으로 설명될 수 있다고 생각한다.

**- 기록되지 않은 세계 (Unrecorded World):**

우리가 AI와 경쟁하며 간과하는 사실은, 세상의 정보 중 텍스트나 숫자로 기록되는 것은 극히 일부라는 점이다. 일상에서 느끼는, 클라이언트와 대화할 때 그의 흔들리는 동공, 회의실의 무거운 공기, 칠판에 휘갈겨 쓴 다급한 글씨 등 이러한 비정형 데이터가 빼곡히 우리 일상을 메운다. AI는 수십조 개의 토큰을 학습하지만, 이러한 '공기(Atmosphere)'를 학습할 수는 없다.

그리고 인간의 '경험'이라는 것은 시간의 축적으로만 볼 게 아니라, 정의되지 않은 문제를 해결하는 과정에서, 위와 같은 '공기'들이 축적된 암묵지(Tacit Knowledge)의 총량이라 볼 수 있다. 젠슨 황 같은 CEO들은 세상을 머리로만 아는 게 아니라, 감각적으로 체화한 것이다. 공기와 암묵적 지식들은 텍스트로 변환되는 순간 정보의 90%가 손실되는, 오직 인간만이 접근 가능한 데이터 영역이다.

**- Chain of Frame: 고차원 현실의 손실 압축:**

인간의 뇌는 이 방대한 비정형 데이터를 어떻게 처리할까? 젠슨 황과 같은 리더는 오감과 감정을 총동원해 초고차원(High-dimensional)의 현실을 '하나의 프레임(Frame)'으로 캡처한다고 나는 생각한다. 그러나 인간의 생물학적 뇌는 용량의 한계가 있기에, 무한에 가까운 현실의 정보를 인간이 수용할 수 있는 차원으로 근사화(Approximation)하여 저장해야 한다. 즉, 뇌가 저장하는 것은 원본 비디오가 아니라, 가장 중요한 특징(Feature)들만 추출된 압축 데이터라고 생각한다.

'우리는 우리가 알고 있는 것보다 사실 더 많이 아는 것일 수 있다'는 차원의 이야기다. "설명은 못하겠는데 그냥 이게 답일 거 같아"라고 말하는 것은, 뇌가 N차원의 데이터를 처리하여 최적해를 도출했으나, 이를 1차원인 언어로 변환(압축)하는 과정에서 병목이 발생하여 '느낌'이라는 메타 데이터로 출력한 것이다.

**- 무의식적 매니폴드 탐색 (Manifold Search):**

뇌과학적으로 해마(기억)는 편도체(감정)와 붙어 있다. 과거 실패했을 때의 '등골이 오싹한 공포'나 성공했을 때의 '전율'은 그 기억을 불러오는 강력한 색인(Index) 역할을 한다. 젠슨 황의 뇌 속에는 감정 태그가 붙은 수만 개의 프레임이 벡터(Vector) 형태로 저장되어 있다. CEO가 결정을 내리는 순간, 그는 의식적으로 과거를 회상(Recall)하는 것이 아니다. 그의 뇌에는 무의식 속에서 저장된 수만 개의 'Chain of Frame' 벡터들을 고차원 공간 상에서 탐색하게 된다.

즉, 젠슨 황의 뇌는 현재의 상황 벡터(시장의 분위기, 현장에서의 엔지지어들의 뉘앙스 등)가 과거 성공했던 그날의 벡터와 유사함을 감지하면, 논리적 단계(A→B→C)를 건너뛰고 결론(A→Z)으로 직행한다. 이는 고차원 공간을 접어서 이동하는 웜홀과 같다. 논리적으로 설명하지 못하지만 정답을 맞힌다는 사실 자체가, CEO들을 포함한 인간 권리자의 뇌가 언어의 차원보다 훨씬 높은 차원에서 데이터를 처리하고 있음을 방증한다.

**– 이것이 왜 인간 리더의 특권인가?:**

AI는 주어진 데이터(Ground Truth)를 기반으로 최적값을 찾지만 젠슨 황은 기록되지 않은 무수히 많은 요소를 통합하여 'AI 시대의 도래'라는 새로운 Ground Truth를 창조했다.

AI가 엑셀 데이터(기록된 세계)를 완벽하게 장악한 2025년에도, 인간 리더들의 뇌 속에 구축된 이 '거대한 비정형 데이터베이스'와 그것을 처리하는 신경망은 AI가 흉내 낼 수 없는 영역이다. 개인의 경험은 파편적이지만, 그 경험이 축적되어 형성된 '나만의 맥락'은 우주에 하나뿐인 고유한(Unique) 데이터셋이기 때문이다. AI도 모르고 나만 아는 지식, 나만 해본 경험, 그래서 펼쳐 낼 수 있는 직관이야말로 우리 개개인들이 가지는 필살기다.

## 6. 2026년을 위한 생존 전략: 안목(Taste)이 권력이다

### (1) 기업의 전략: 초미세 벤치마크와 운영의 혁신

AI 도입이 실패하는 이유는 기술이 부족해서가 아니라, 기업의 '마인드셋'과 '구조'가 준비되지 않았기 때문이다.

**- 초미세 벤치마크(Custom Evals):**

기업 입장에서는 AI 모델들의 MMLU 점수나 코딩 테스트 점수가 딱히 의미가 없다. "우리 회사의 M&A 자문 계약 문서의 초안을 작성할 때, 20년 차 파트너 변호사만큼 독소 조항을 잘 걸러내는가?"와 같은 커스터마이징이 된 초미세 벤치마크가 수천 개 필요하다. 업계 사람들이 딱 필요로 하는 "찐벤치마크"를 만들어야 한다는 이야기다. 각 회사들은 해당 도메인의 '성공 기준'을 먼저 정의하고 선점해야 한다.

**- 운영자(Operator) 중심:**

따라서, AI 프로젝트는 IT 부서가 아닌, 현장의 고통(Pain point)을 가장 잘 아는 최고의 관리/운영자가 주도해야 한다. 2~3개의 핵심 과제(예: 결제 업무 자동화 20% 향상 등)에 집중하고, 성공 시 파격적인 보상을 걸어야 한다.

**- 데이터 미니멀리즘:**

회사가 가진 모든 데이터를 '데이터 레이크'에 모으고 정제하는 것은 시간이 3~5년은 걸리는 일이라고 한다. 소수의 핵심 과제 해결에 필요한 'Use Case' 중심의 데이터 5~6개 변수만 빠르게 통합하는 게 답일 수 있다. 이와 관련해, Swiss Gear라는 회사는 750개의 데이터 테이블을 통합하는 것을 포기하고, 재고와 관련한 핵심 변수만 빠르게 통합했다. 그 결과, 2개월 만에 재고 예측 시스템을 구축한 사례가 있다.

### (2) 개인의 진화: 초월적 T자형 인재 (One-Person Unicorn)

실행이 병목이 아니고 진짜 병목이 Taste라면 개인에게 요구되는 인재상은 명확해진다.

**- One-Person Unicorn:**

AI라는 무한한 실행 자원을 비서로 두고, 혼자서 기획-디자인-개발-마케팅의 전 과정을 관통하는 초월적 'T자형 인재'가 되어야 한다. T자형 인재에서, 가로축은 AI를 통해 무한히 확장될 수 있고, 세로축은 기술적 깊이라기 보다는 '무엇이 아름다운가에 대한 취향', '실패를 두려워하지 않는 배짱'으로 정의된다. 가로축은 AI를 본인의 과업 모든 것에 연관해보면서 결과물의 가치를 증강하는 것으로 갈음할 수 있지만, 물론, 세로축의 taste라는 것은 빠르게 얻어낼 수 있는 것은 전혀 아니다.

**- 고통을 통한 뇌의 튜닝:**

고유의 taste를 갖춘 전문가가 되기 위해서는 편안함을 거부해야 한다. '워라밸'의 가치는 잠시 놓아주고 스트레스를 받으면서 몰입해야 한다. AI에게 실행을 맡기고 결과만 받아먹는 사람은 뇌의 가소성을 잃고 도태될 확률이 높기 때문이다. 이와 관련해 서울대학교 뇌인지과학과의 이인아 교수는 "AI와 1,000시간 대화한다고 전문성이 생기지 않는다"고 단언했다. 인간의 전문성이라는 것은 현장에서 뒹굴고, 깨지고, 수십일을 밤새우는 고통스러운 실행 과정에서 형성된 신경망의 총체라는 것이다.

### (3) Final Insight: 피드백 루프의 설계자

AI가 1초 만에 100개의 시안을 내놓을 때, "이것이 정답이다"라고 고를 수 있는 압도적인 안목은 오직 자신의 뇌에 'Chain of Frame'을 고통스럽게 새긴 자만이 가질 수 있다. AI를 도구로 쓰되 더 깊고 고통스러운 현장의 문제 해결에 뛰어들어 대체 불가능한 안목을 길러야 한다.

우리가 해야 할 일은 'AI에게 일을 시키고 빠지는 것'이 아니다. AI가 해온 일을 내 눈으로 확인하고, 내 노하우(암묵지)를 더해 다시 AI를 가르치는 '피드백 루프(Feedback Loop)'를 설계하는 것이다. 이 루프를 얼마나 정교하게 설계하느냐가 승패를 가를 것이다. 기계가 모든 것을 수행하는 세상에서, 인간이 노예가 아닌 주인으로 남을 수 있는 유일한 길은 기록되지 않은 세계를 읽어내는 야생의 감각을 유지하는 것이다.

# II. 엔비디아의 Groq 자산인수

## 1. 구조적 병목과 규제 우회

### (1) Groq 자산인수에 200억 달러나 왜 썼나? '시간'을 매입하다

2025년 크리스마스 이브에 발표된 엔비디아와 Groq의 200억 달러 규모의 딜은 그 타이밍만큼이나 거래 구조가 교묘했다. 전액 현금으로 치러진 이 거래의 명목은 인수합병(M&A)이 아닌 '자산 매입 및 인력 채용(Asset Purchase & Acquihire)'이다. 이는 미 법무부(DOJ)와 FTC가 엔비디아의 독주를 막기 위해 눈에 불을 켜고 있는 상황에서, 규제 당국의 감시망을 무력화하기 위해 설계된 것이라 할 수 있다.

이로써 형식상 Groq이라는 이름의 법인은 사라지지 않고 CFO가 대표 자리를 대체한다. 클라우드 서비스 기업으로 명맥을 유지하지만, 물론 껍데기에 불과한 듯 보인다. 엔비디아가 실제로 가져간 것은 창업자 조나단 로스(Jonathan Ross)를 포함한 핵심 엔지니어링 팀과 LPU(Language Processing Unit) 아키텍처 IP 전체이기 때문이다. 시장은 이를 두고 엔비디아가 경쟁사를 통째로 삼키는 리스크를 피하면서도, 실질적으로는 경쟁사의 뇌수(Brain)와 심장(IP)을 모두 가져가 버린 '위장된 인수'로 평가한다.

이 딜의 본질은 AI 가속기 부문의 경쟁자 제거에만 있지 않다. 우리는 엔비디아가 물리적 시간을 산 것에 크지만 합리적인 비용을 치렀다고 생각한다. GPU의 가장 큰 경쟁자인 구글 TPU를 설계했던 팀의 유전자는 물론이고, 그들이 10년간 갈고닦은 '결정론적(Deterministic) 설계' 노하우를 즉시 확보함으로써, 자체 로드맵을 최소 2년 이상 단축시킨 것으로 풀이되기 때문이다. 이는 하이퍼스케일러들이 자체 칩(ASIC)으로 이탈하려는 움직임을 또 한 번 봉쇄할 수 있고, 향후 5년의 AI 인프라 패권을 결정짓는 엔지니어링의 시발점일 수 있다.

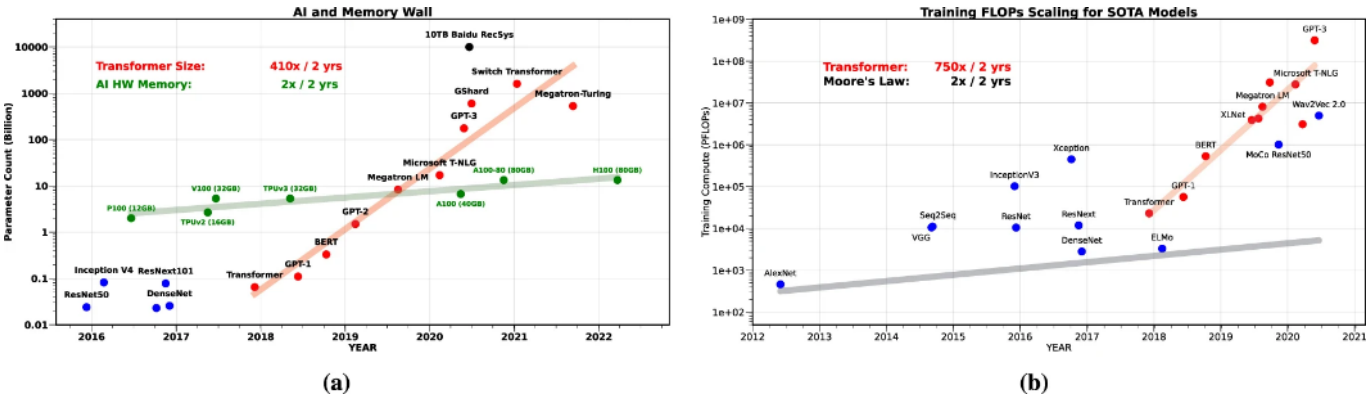
### (2) 하드웨어의 절규: 산술 강도(Arithmetic Intensity)의 불일치

현재 AI 인프라가 겪고 있는 고통은, 모델의 크기와 추론 복잡도는 기하급수적으로 증가하는데, 메모리 대역폭은 선형적으로 가하는 데 그치는 구조적 병목 때문이다. 엔비디아가 Groq을 필요로 했던 기술적 이유는 H100이나 B200과 같은 기존 GPU 아키텍처가 가진 태생적 한계, 즉 '산술 강도의 불일치'에서 찾을 수 있다.

LLM의 추론, 특히 디코딩(Decode) 단계는 본질적으로 순차적이다. 다음 토큰 하나를 생성하기 위해 수천억 파라미터 모델의 모든 가중치를 HBM에서 불러와야 하지만, 이때 정작 수행하는 연산은 바이트당 고작 1~2 FLOPs 수준이다.

결과적으로 GPU의 텐서 코어는 데이터가 HBM에서 도착하기를 기다리며 사이클의 대부분 시간 동안 '놀고' 있다. GPU 구매자들이 지불하는 천문학적인 비용 대부분은 이 '노는 연산기'에 해당하기 때문에, 이는 매우 자본 비효율적이다. 지난 수년간 연산 속도(FLOPS)는 2년마다 750배 증가했지만, 대역폭은 1.6배 증가하는 데 그친 'Memory Wall'이 만든 참사다.

**그림 7. 트랜스포머 모델의 크기(좌측)와 연산량(우측)이 2년마다 수백 배씩 폭증할 때, 하드웨어 메모리의 발전 속도는 고작 2배에 그침**




자료: IEEE Micro "AI and Memory Wall", 미래에셋증권 리서치센터

### (3) KV Cache의 기하급수적 폭발과 물리적 파탄

문제는 여기서 그치지 않는다. 모델이 단순한 문답을 넘어 'Reasoning(추론)'을 수행하기 시작하면서 Context Window가 100M(1억) 토큰 단위로 확장되고 있다. 이에 따라 KV Cache(Key-Value Cache)의 용량은 HBM이 감당할 수 없는 수준으로 폭증했다. 이를 DeepSeek V3 모델(MoE)을 기준으로 계산해보면 상황의 심각성이 드러난다.

** KV Cache 용량 공식: 2 *2 * head_dim * n_heads * n_layers * seq_length * batch_size*

DeepSeek V3에서는 토큰당 약 34.3KB의 메모리가 필요하다. 100M Context를 가진 Reasoning 모델을 구동하려면, 사용자 단 한 명을 위해 3.2TB의 메모리가 필요하다는 말이 된다. 현재 가장 진보된 B200 GPU(192GB HBM)를 사용하더라도, 3.2TB를 확보하려면 GPU 17장이 필요하다. 사용자 한 명의 질문을 처리하기 위해 5억 원 이상의 하드웨어를 할당해야 한다는 계산이 나온다.

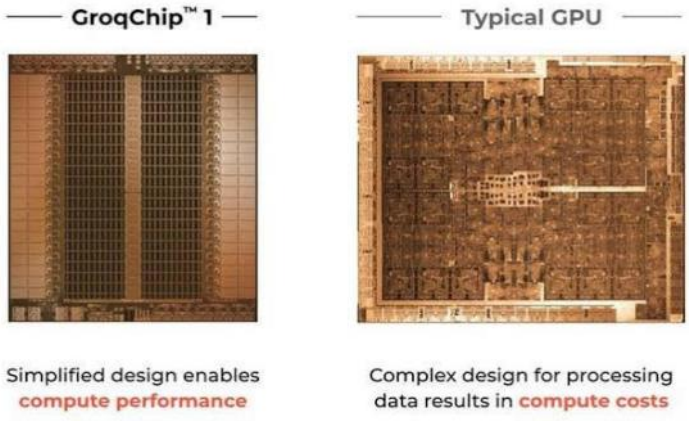
일론 머스크가 말한 "Amazing Abundance"는 추론 비용이 한계비용 제로에 수렴할 때만 가능하다. 그러나 현재의 GPU 아키텍처로는 HBM 용량과 대역폭의 물리적 한계 때문에 놀라운 풍요를 달성하는 것은 요원해 보였다. 엔비디아는 이 물리적 장벽을 정면 돌파하기 위해서라도 다른 방식의 아키텍처가 절실했을 수 있다.

## 2. Feynman 아키텍처: 이종 적층(Heterogeneous Integration)

### (1) SRAM 스케일링의 종말과 모놀리식의 한계

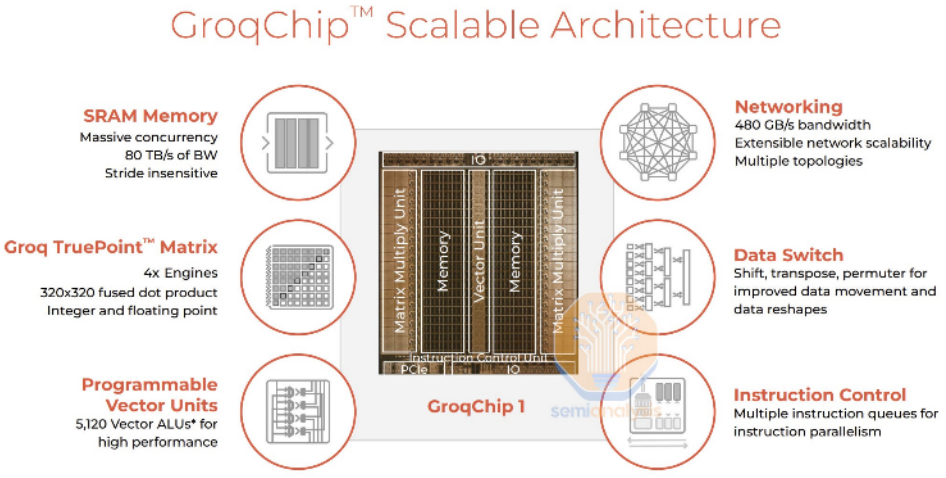
Groq이 결국 독자 생존을 포기하고 엔비디아 제국의 일부로 흡수된 결정적인 원인은, 그들의 기술력이 부족해서가 아니라 "물리적 공정의 배신" 때문이었다. Groq의 아키텍처는 거대한 단일 칩(Monolithic Die)에 SRAM을 최대한 많이 때려 박는 구조다.

**그림 8. Groq(좌측)은 복잡한 스케줄러와 캐시 컨트롤러를 제거한 VLIW 아키텍처**
**다이 면적 대부분을 연산(ALU)과 SRAM에 할당해 밀도를 극대화**
**반면 GPU(우측)는 런타임 동적 스케줄링을 위한 제어 로직이 복잡하게 얽혀 있음**
**실제 연산기가 차지하는 유효 면적 효율이 상대적으로 떨어짐.**




자료: Groq, 미래에셋증권 리서치센터

**그림 9. HBM 없이 칩 내부에 분산된 230MB SRAM이 초당 80TB라는 경이로운 대역폭을 제공**
**하지만 HBM 컨트롤러 같은 외부 메모리 인터페이스의 부재가 결국 한계가 되었음**
**LLM 구동 시 수천 개의 칩을 연결해야 하는 '확장성(Capacity)의 재앙'을 초래**




자료: Groq, SemiAnalysis, 미래에셋증권 리서치센터

그런데, 이 방식은 파운드리 공정의 한계 앞에서 더 이상 유효하지 않게 되었다. TSMC의 공정 로드맵을 들여다보면, 로직(Logic) 트랜지스터는 N5에서 N3E, N2로 가면서 계속 작아지지만, SRAM 비트셀의 면적 축소는 사실상 멈추었기 때문이다.

N2 공정 기준 SRAM 비트셀 크기는 약 0.0175μm² 수준에서 정체되어, N3 대비 유의미한 밀도 개선이 없다. 반면, 선단 공정(Advanced Node)의 웨이퍼 가격은 세대가 바뀔 때마다 폭등하고 있다. 비싼 3nm, 2nm 웨이퍼 면적의 절반 이상을, 크기가 줄어들지도 않는 구형 기술인 SRAM으로 채우는 것은 비용 효율 측면에서 재앙이다.

Groq이 최신 공정을 썼음에도 칩 하나에 고작 230MB의 SRAM밖에 넣지 못했고, Llama 3 70B 같은 거대 모델을 돌리기 위해 칩 576개를 연결해야 했던 비효율의 근본 원인이 바로 여기에 있다. 칩 내부의 속도는 빛처럼 빨랐지만, SRAM 용량 부족으로 인해 칩을 수백 개 연결해야 했고, 결국 칩과 칩 사이의 네트워크 비용이 배보다 배꼽이 더 커지는 구조적 모순에 빠진 것이다. 결국에는 엔비디아도 Groq도 메모리가 문제인 것이다.

**표 3. 2024년 1분기 기준, AI 추론 인프라 경제성 및 성능 비교 (H100 vs Groq 8-Rack)**

| | 세부 항목 | H100 (지연시간 최적화) | H100 (처리량 최적화) | Groq LPU 시스템 (8개 랙) | 비교 (전략적 함의) |
|---|---|---|---|---|---|
| 1. 인프라 구성 | 추론 유닛당 칩 수 | 8개 | 2개 | 576개 | Groq은 동일 모델 구동을 위해 칩이 훨씬 더 필요 (SRAM 용량) |
| | 시스템 물리적 규모 | 1개 서버 (4U) | 1/4개 서버 | 8개 랙 (Rack) | 공간 비용율: 데이터센터 상면(Space) 점유율이 압도적으로 높음 |
| | 호스트 CPU 필요량 | 2개 (서버당) | 0.5개 (공유) | 144개 | 불필요한 범용 하드웨어(CPU/RAM) 중복 투자 발생 |
| 2. 비용 구조 | 칩 개당 원가 (추정) | $3,200 (고가) | $3,200 (고가) | $650 (저가) | 칩 자체는 싸지만, 시스템 레벨에서는 의미가 없음 |
| | 추론 유닛당 BOM 비용 | ~$29,600 | ~$7,400 | ~$604,800 | 비용 폭탄: Groq 시스템 구축 비용이 H100 대비 20~80배 높음. 시스템을 구성하는 데 드는 비용(CPU 144개, 8개 랙 공간, 전력)이 배보다 배꼽이 더 큰 구조<br>운영비 제약: 전력 소모와 상면 비용이 너무 커서 지속 불가능 |
| | 월간 호스팅 OpEx 비용 (전력/쿨링 등) | ~$2,586 | ~$646 | ~$58,409 | |
| | 월간 총비용 (CapEx+OpEx) | $11,474 | $2,868 | $122,400 | Groq은 H100 대비 월 유지비가 10배 이상 소요됨 |
| 3. 성능 지표 | 배치 사이즈 | 2 (작음) | 64 (큼) | 3 (매우 작음) | SRAM의 한계: Groq은 메모리 부족으로 동시 처리량(Batch)을 늘릴 수 없음 |
| | 동시 접속자 수 | 2명 | 64명 | 48명 | HBM 기반 GPU가 대규모 트래픽 처리에 훨씬 유리함 |
| | 초당 처리 토큰 (TPS) | 840 | 1,920 | 24,000 | Groq의 속도는 압도적이나, 이는 '물량 공세(576개)'의 결과임 |
| 4. 경제성 분석 | 100만 토큰당 원가 | $5.20 | $0.57 | $1.94 (원가 기준) | 처리량 최적화된 H100이 가장 저렴함 (압도적 가성비),<br>Groq은 현재 구조에서 수익을 낼 수 없음 (적자 구조 고착화) |

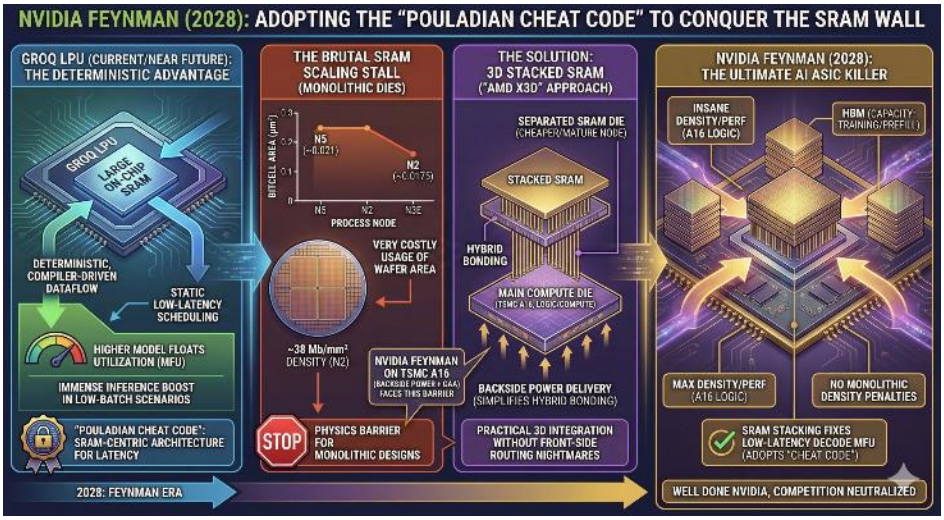
자료: Groq, SemiAnalysis, 미래에셋증권 리서치센터

### (2) 엔비디아의 해법: 3D 적층(Stacking)과 하이브리드 본딩

엔비디아의 차세대 아키텍처인 "Rubin(내년 본격 출하)" 이후에 등장할 "Feynman(파인만, 2028년 출시 예정)"은 Groq과의 결합을 수평이 아닌 수직으로 분리하는 이종 집적(Heterogeneous Integration)으로 구현할 것으로 관측된다.

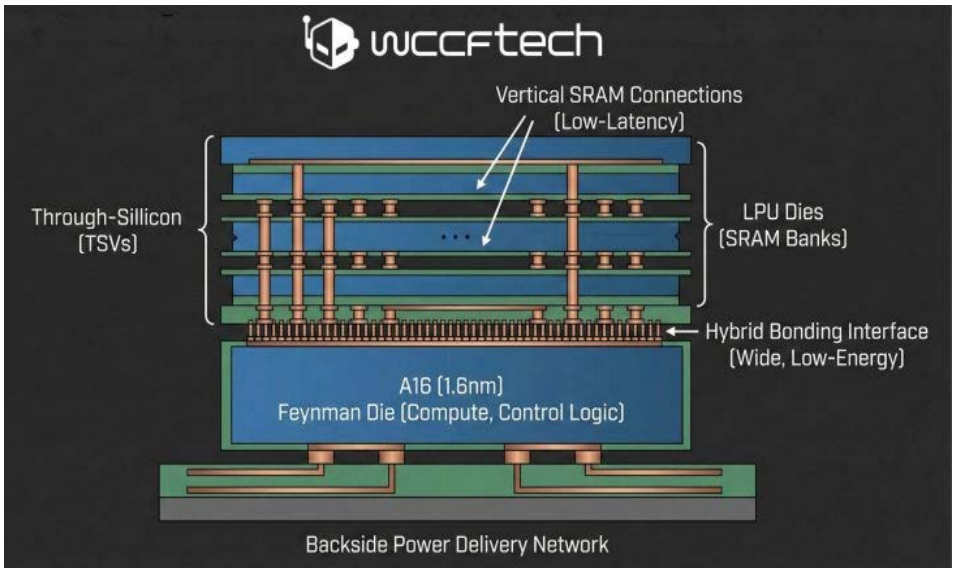
유명한 반도체 엔지니어인 Ben Pouladian에 따르면, 엔비디아는 Groq의 SRAM을 더 이상 비싼 평면에 배치하지 않고, 연산 로직 위에 빌딩처럼 쌓아 올리는 방식을 구현할 것으로 사료된다.

**그림 10. 좌측의 Groq LPU는 SRAM을 평면에 배치하여 빠른 속도를 냈지만,**
**공정이 미세화되어도 SRAM 크기가 줄어들지 않는 '스케일링 정체'에 부딪힘.**
**중앙 그래프에서 TSMC의 N5에서 N2 공정으로 가더라도 SRAM 밀도는 거의 제자리걸음.**
**엔비디아 차세대 칩 Feynman(우측)은 이를 극복하기 위해,**
**컴퓨팅 다이 위에 저렴한 공정으로 만든 SRAM 다이를 수직으로 쌓는 '3D Stacking'을 도입할 것.**
**이때, TSMC A16 공정의 후면 전력 공급(BSPDN) 기술이 전면부 배선 공간을 확보해 주어,**
**이러한 고밀도 이종 칩 적층(하이브리드 본딩)이 가능해진 것으로 풀이.**
**이로써 Feynman은 HBM의 '무한한 용량'과 적층형 SRAM의 '즉각적 반응 속도'를 동시에 확보.**




자료: 나노 바나나 프로, Ben Pouladian, 미래에셋증권 리서치센터

**그림 11. Feynman의 예상 구조, TSMC A16 공정으로 제작된 고성능 컴퓨팅 및 Base Die이 위치.**
**상단에는 Groq의 기술이 적용된 SRAM Banks가 여러 층으로 적층돼 초저지연 메모리 풀을 형성.**
**두 영역은 초미세 하이브리드 본딩으로 접합되고, 칩을 관통하는 수직 연결(TSV) 통로들이 층층이**
**배치되어, 하단의 연산기가 상단의 수많은 SRAM 뱅크에 즉각적으로 접근할 수 있게 설계.**
**최하단의 후면 전력 공급망(Backside Power Delivery)은 전력을 뒤로 빼주어,**
**칩 전면 공간을 온전히 이러한 복잡한 3D 신호 연결에 쓸 수 있게 조성.**




자료: wccftech, 미래에셋증권 리서치센터

**하단 (Base Die – The Brain):**

TSMC의 최선단 A16 공정을 사용한다. 여기에는 면적 대비 성능 효율이 극도로 중요한 텐서 코어(연산기)와 복잡한 제어 로직, 그리고 HBM 컨트롤러만 남긴다. 비싼 땅에는 가장 비싼 건물만 짓는 원리다.

**상단 (SRAM Die – The Memory):**

최선단 공정이 아닌, 값싸고 수율이 완벽에 가까운 레거시(성숙) 공정을 사용한다. 여기에 Groq의 기술이 적용된 대용량 SRAM 전용 다이를 만든다. SRAM은 미세 공정 혜택을 못 받으니, 굳이 비싼 공정을 쓸 이유가 없다.

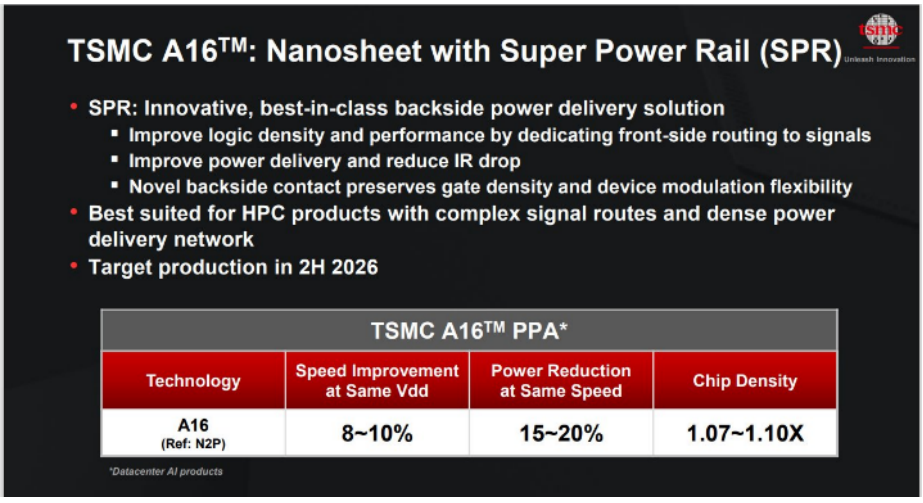
**초미세 연결 (The Glue):**

TSMC의 SoIC 기술(Hybrid Bonding)을 이용해 두 다이를 수직으로 접합한다. 기존의 범프(Bump) 연결이 아니라 구리와 구리를 직접 붙이는 것이다. 이렇게 하면 상단 SRAM과 하단 로직 사이의 물리적 거리가 사실상 '0'에 수렴하기 때문에, 대역폭 병목 없이 마치 CPU 바로 옆에 붙어 있는 L1 캐시 수준의 속도로 거대 메모리 풀에 접근할 수 있게 된다.

### (3) TSMC의 A16 공정과 후면 전력 공급(BSPDN)의 필연성

이때 중요해지는 것은 TSMC의 A16 공정의 성공과 BSPDN이 된다. 기존 칩(Front-side Power)은 웨이퍼 전면에 전력 공급선과 데이터 신호선이 함께 복잡하게 얽혀 있다. 이 상태에서 상단에 또 다른 칩(SRAM)을 붙이고 수직으로 구멍(TSV)을 뚫으려 하면, 전력선이 방해되어 신호를 연결할 공간이 턱없이 부족해진다.

**그림 12. Feynman 칩 생산의 기반이 될 TSMC A16 공정의 핵심인 SPR은, 내년 하반기 양산 목표**
**SPR은 칩의 로직 밀도를 높이고 전압 강하를 방지하여, 연산 성능과 전력 효율을 15~20% 개선.**
**특히 복잡한 신호 라우팅이 필수적인 3D 적층형 칩에 최적화되어 있어, 복층 구조를 실현하는 열쇠**




자료: TSMC, 미래에셋증권 리서치센터

이때 TSMC A16 공정의 슈퍼 파워 레일(Super Power Rail, SPR) 기술은 전력망을 웨이퍼 뒷면으로 통째로 옮겨버린다(BSPDN 기술). 이로써, 전면부는 온전히 데이터 통신을 위한 신호선과, 상단 SRAM 다이와 연결될 고밀도 수직 연결(TSV)을 위한 고속도로로 활용될 수 있다. 즉, A16 공정은 3D 적층을 위한 물리적 공간을 확보해 주는 전제 조건이다.

그런데 고밀도 연산 다이 위에 메모리 다이를 이불처럼 덮으면, 하단에서 발생한 열이 빠져나가지 못하고 갇히는 '열섬 현상'이 발생할 수 있는데, BSPDN은 굵직한 전력선이 뒤쪽으로 빠지면서 하단부의 열 배출 경로를 확보하는 데 결정적인 역할을 한다.

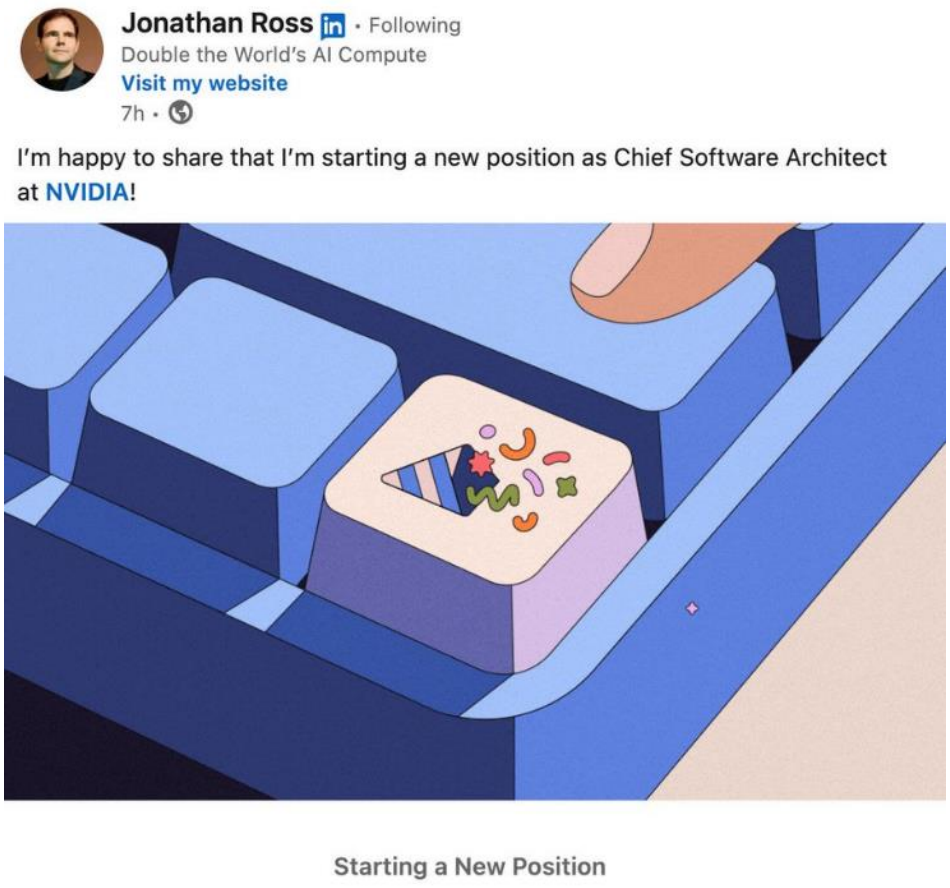
결국 Feynman은 "HBM의 압도적인 용량(Capacity)"과 "적층형 SRAM의 즉각적인 반응 속도(Latency)"를 단일 칩 안에서 동시에 구현하는 칩이 될 것으로 보인다. 엔비디아는 Groq의 IP를 이 3D 적층 구조의 상단부에 이식함으로써, 반도체 업계가 수십 년간 통곡의 벽으로 여겼던 '메모리 월'을 패키징 기술로 우회하여 정복할 수도 있다. 이것이 엔비디아가 200억 달러를 지불하며 얻어낸 '물리적 시간'의 정체다.

## 3. 소프트웨어 해자: 결정론적 통제와 물리적 시간의 동기화

### (1) 엔비디아가 진짜 산 것: 하드웨어가 아닌 '결정론적(Deterministic) 두뇌'

Cerebras와 같은 peer 그룹의 경쟁사들은 Groq의 성능을 두고 "단지 SRAM을 많이 써서 빠른 것일 뿐"이라며 그 가치를 평가절하해왔다. 하지만 이는 반만 맞고 반은 틀린, 경쟁사의 불안감이 투영된 주장으로 생각한다. 엔비디아가 200억 달러를 주고 인수한 것은 SRAM 하드웨어만이 아니다. 물리적 메모리가 아니라, 그 메모리를 시간의 오차 없이 제어하는 소프트웨어(컴파일러)와 아키텍트들을 산 것이다.

**그림 13. Groq의 CEO가 본인 링크드인에 남긴 말 "저는 이제 엔비디아에서 최고 SW 담당자에요"**
**엔비디아가 Groq에게 원한 것은 하드웨어가 아니라 소프트웨어라는 점을 주장을 할 수 있는 증거물**

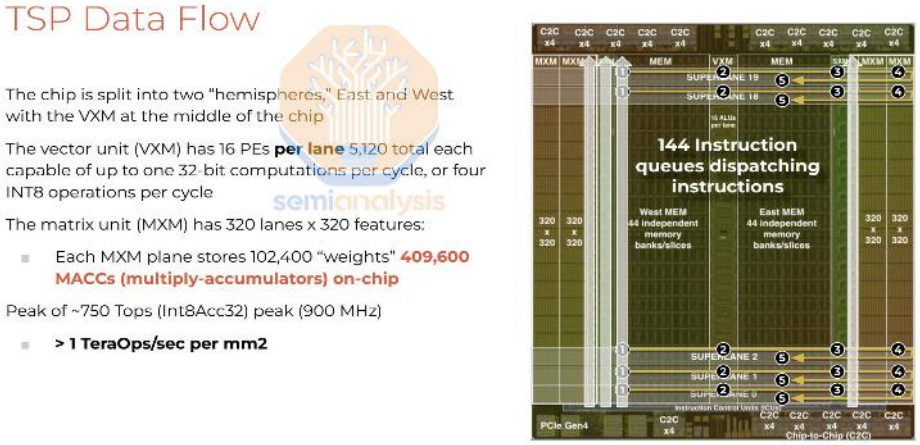



자료: Janathan Ross, 미래에셋증권 리서치센터

일반적인 GPU 아키텍처(CUDA)는 본질적으로 '비결정론적(Non-deterministic)'이다. 실행 런타임에 데이터가 캐시 메모리에 있는지 확인하고(Cache Hit/Miss), 없으면 가져오고, 그 과정에서 스레드 순서를 조정하느라 끊임없이 판단한다. 이 과정에서 필연적으로 예측 불가능한 지연(Jitter)과 오버헤드가 발생한다.

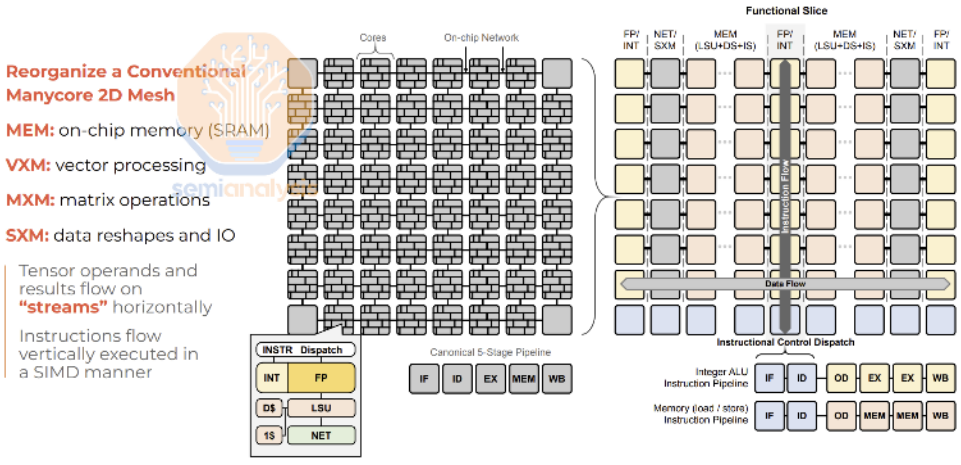
반면, Groq의 핵심 자산인 '결정론적 컴파일러'는 데이터가 언제, 어느 SRAM 뱅크에서 나와서, 어느 연산기로 이동할지를 컴파일 단계에서부터 나노초(ns) 단위로 '미리 계산'한다. 런타임에는 하드웨어가 어떠한 판단도 하지 않고, 정해진 악보대로만 연주하면 된다.

**그림 14. 144개의 명령어 큐가 VLIW 방식으로 320개의 Superlanes에 동시에 명령을 내리는 구조.**
**중앙의 연산 유닛(VXM/MXM)은 SRAM에서 데이터를 즉시 공급받아(80TB/s),**
**데이터 로딩을 기다리는 유휴 시간(Idle Time)이 없다.**
**이러한 정교한 데이터 흐름 제어 기술은 엔비디아가 인수한 'Groq 컴파일러 팀'의 핵심 역량이며,**
**향후 CUDA 생태계의 비밀 병기가 될 것.**




자료: Groq, SemiAnalysis, 미래에셋증권 리서치센터

**그림 15. 전통적인 2D 메쉬 구조와 다른 Groq(우측) 칩의 데이터 흐름**
**기능별(메모리, 벡터, 행렬)로 유닛을 길게 슬라이스하여 데이터 흐름을 일방통행으로 만듦.**
**데이터는 수평(East-West)으로 흐르고 명령어는 수직(North-South)으로 흐르는 구조.**
**이로써 데이터 충돌 없는 파이프라인 처리가 가능.**
**이 구조는 컴파일러가 데이터의 위치와 이동 타이밍을 100% 예측할 수 있게 해주어,**
**하드웨어 레벨의 중재(Arbitration) 오버헤드를 제거.**




자료: Groq, SemiAnalysis, 미래에셋증권 리서치센터

이 기술이 왜 엔비디아에게 필요할까? Feynman 칩의 상단에 적층될 거대한 SRAM 풀(Pool)을 100% 효율로 제어할 유일한 방법이기 때문이다. 수천 개의 뱅크로 쪼개진 SRAM을 마치 CPU 바로 옆의 L1 캐시처럼 지연 없이 쓰게 만드는 것은, 물리적 연결만으로는 불가능하다. 하드웨어의 복잡성을 소프트웨어가 모두 떠안아, 실행 시점의 불확실성을 '0'으로 만드는 소프트웨어 기술이 필수적이다. 엔비디아는 이 소프트웨어 두뇌를 흡수함으로써, 향후 3D로 적층할 대용량 SRAM 다이를 통제할 열쇠를 손에 쥐게 됐다.

### (2) 빛의 속도를 극복하는 기술: 준동기식 클럭(Plesiochronous Clock)

이 딜의 진짜 무서움은 칩 내부를 넘어 데이터센터 전체를 장악하려는 야심에 있다. 바로 Groq이 보유한 '준동기식 클럭(Plesiochronous Clock)' 기술이다. 경쟁사인 Cerebras나 SambaNova 역시 SRAM 중심 아키텍처를 표방하지만, 이들은 랙(Rack)과 랙 사이의 통신에서 여전히 비동기적 요소를 가지고 있다.

보통 서버 랙(Rack)과 랙 사이의 통신은 '비동기(Asynchronous)'다. 물리적 거리가 멀고 각 서버의 시계가 미세하게 다르기 때문에, 데이터를 보낼 때마다 "받을 준비 됐니?(Handshake)", "응, 보내(Ack)" 하는 확인 절차가 필수적이다. 빛의 속도가 아무리 빨라도 이 절차 때문에 병목이 생긴다. 이는 데이터센터 확장의 가장 큰 걸림돌이다.

하지만 Groq은 데이터센터 스케일에서 '준동기식 클럭'을 쓰는 것이다. 이는 랙 전체에 정밀한 기준 클럭 신호를 뿌리고, 칩들이 실시간으로 자신의 시계 오차를 보정하게 만드는 기술이다. 게다가 Groq의 컴파일러는 케이블의 길이와 물리적 거리에 따른 지연 시간을 이미 '상수'로 알고 있기 때문에 칩 A가 데이터를 쏘면, 칩 B는 확인 절차 따위 없이 정확히 계산된 그 시간에 손을 뻗어 데이터를 잡게 된다. 이 기술 덕분에 물리적으로 분산된 수많은 SRAM 뱅크들을 단일 칩 내부의 공유 메모리처럼 지연 없이 활용할 수 있는 것이다.

엔비디아는 이러한 스케일링 기술을 NVLink 도메인에 적용할 것이다. 랙과 랙 사이를 넘나드는 수천 개의 AI 가속기가 마치 하나의 거대한 칩처럼 미세한 오차 범위 내에서 동기화되어 돌아간다면 어떻게 될까? 이는 경쟁사들이 흉내 내기 힘든 물리적 해자가 된다.

### (3) cuTile: 개발자를 위한 마법의 추상화

엔비디아의 전략적 치밀함은 이 복잡하고 어려운 기술을 개발자들에게는 효과적으로 숨긴다는 데 있다. 엔비디아는 CUDA 13.1에서 `cuTile`이라는 새로운 추상화 계층을 조용히 도입한 적이 있다.

기존 Groq 아키텍처의 컴파일러의 성능이 괄목할 것만큼이나, "개발자가 메모리 위치와 타이밍을 일일이 지정해야 한다"는 고난도 진입 장벽이 문제였다. 하지만 엔비디아가 제공하는 cuTile을 사용하면, 개발자는 엔비디아 GPU든 Groq의 LPU든 하드웨어의 복잡한 타일 구조를 몰라도 된다. 그저 CUDA로 코드를 짜면, 컴파일러가 알아서 작업을 배분할 수 있다. 순차적이고 반복적인 디코딩 작업은 LPU 타일(결정론적 영역)로, 유연성이 필요한 작업은 GPU 코어(동적 영역)로 보낼 수 있다. 즉, "Groq은 컴파일러가 복잡하다"는 약점을, 엔비디아는 "그냥 CUDA 쓰세요, 알아서 LPU로 보내줍니다"라는 한 마디로 무력화시킨다.

## 4. 메모리 계층의 재편과 경제성의 승리

### (1) 16-Hi HBM 전쟁: 1,000억 달러 시장을 향한 극한의 레이스

마이크론의 최신 전망에 따르면, HBM 시장 규모는 2025년 350억 달러에서 2028년 1,000억 달러로 단 3년 만에 3배(285%) 성장할 것으로 예측된다. 이는 당초 예상보다 목표 달성 시기가 2년이나 앞당겨진 수치다. 이 파도 위에서 엔비디아는 파트너사들에게 극한의 제조 능력을 요구하고 있다.

엔비디아가 2026년 하반기 공급을 요구한 16단(16-Hi) HBM4는 JEDEC 표준 높이(775μm)를 맞추기 위해 웨이퍼를 30μm 두께로 깎아내야 하는 말 그대로 제조의 물리학을 시험하는 전쟁터다. 물론 엔비디아가 막대한 요구를 하는 것만큼 HBM의 중요성을 대변하기도 한다.

### (2) Storage Next: 저장에서 연산으로, SSD의 신분 상승

수 테라 바이트에 해당하는 KV Cache를 값비싼 HBM에 담는 것은 경제적으로 몹쓸 짓이다. 엔비디아는 이 문제를 해결하기 위해 "Storage Next" 이니셔티브를 꺼내 들었다. 메모리 계층의 최하단에 있던 SSD를 GPU가 직접 부리는 'L4 캐시'로 격상시키는 전략이다. 우리가 여러 자료로써 강조한 계층형 메모리에 대해서 엔비디아가 본격적으로 나선 것이다.

CPU를 거치지 않고(Bypass) SSD가 GPU에 데이터를 직접 쏘는 방식으로써, 2027년까지 메모리 업체들은 1억 IOPS를 달성한다는 목표를 세웠다. NAND가 차가운 데이터 창고가 아니라 지능형 스토리지로 격상되는 것을 시사한다.

이를 위해 SK하이닉스는 2026년 1월, 샌디스크와 협력하여 HBF(High Bandwidth Flash)라는 폼팩터의 알파 버전을 공개할 예정이다. 이는 NAND를 HBM처럼 적층하여 대역폭을 극대화한 것으로 성공시 게임 체인저가 될 것이다. (필자는 HBF 기술에 회의적이다)

한편, CXL 기반 메모리 시스템 값비싼 HBM과 느린 SSD 사이의 간극을 메워주는 거대한 '중간 데이터 풀 지대' 역할을 하며, 엔비디아 서버의 가성비와 효율성을 동시에 최적화할 것으로 풀이된다.

### (3) Groq의 합류: 계층의 완성

여기에 엔비디아가 인수한 Groq의 기술이 더해지면 비로소 '완벽한 계층'이 완성된다. 기존의 “HBM (빠름/작음) - CXL (중간) - SSD (느림/큼)” 구조에서, HBM조차도 해결하지 못했던 '초저지연 디코딩' 영역을 Groq의 SRAM 기술이 담당할 수 있기 때문이다.

- L0/L1: GPU 레지스터 & 캐시 (즉시 연산)
- L2 (New): Feynman 상단의 3D Stacked SRAM (Groq 기술) - 지연 없는 즉각적 반응
- L3: HBM4 - 고대역폭, 모델 가중치 및 Hot Data
- L4: CXL & AI-N P(SSD) - 무한한 용량, Cold Data & Long Context

이 모든 계층은 엔비디아의 단일 소프트웨어 스택(CUDA) 안에서 통합 제어된다. 개발자는 데이터가 어디에 있는지 알 필요가 없다. 그저 "가져오라"고 명령하면, 하드웨어가 알아서 가장 빠른 경로로 데이터를 퍼 나른다. " SRAM으로 속도를, HBM으로 용량을, SSD로 확장을" 갖춘 아키텍처가 오고 있다. 이것이 우리가 엔비디아의 행보에 계속 주목하는 이유다.

# III. AI & Coding

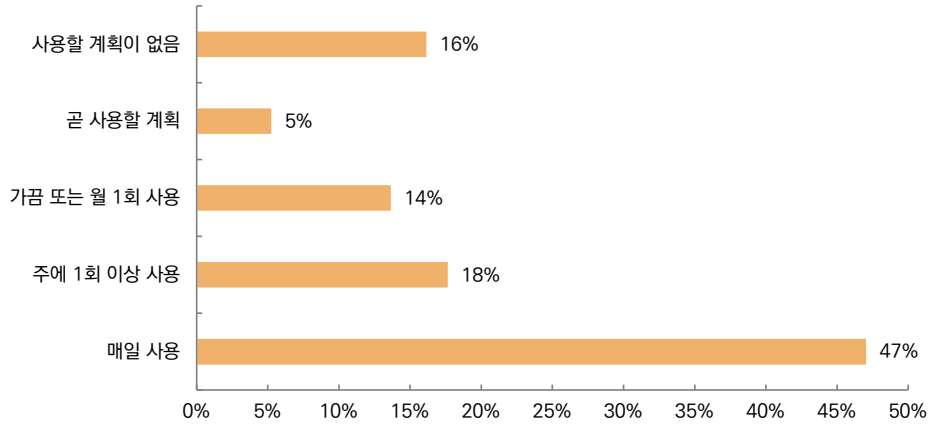
## 1. 확장하는 바이브 코딩

2025년 12월, AI 코딩 도구 시장은 폭발적 성장을 기록했다. 12월 18일, 자연어 기반의 AI 앱 빌더 스타트업인 Lovable은 CapitalG와 Menlo Ventures가 주도한 3억 3천만 달러 규모의 시리즈 B 투자를 발표했다. NVentures, Salesforce Ventures, Databricks 등도 투자자 명단에 이름을 올렸다. 밸류에이션은 66억 달러로, 불과 5개월 전인 7월 18억 달러에서 3.67배 뛰었다. 그리고 가파른 매출 성장이 이러한 평가를 뒷받침했다. 8개월 만에 연간 반복 매출(ARR)이 1억 달러에서 2억 달러로 두 배가 됐고, 일일 생성 프로젝트는 10만 건을 넘어선다.

한편 지난 11월 13일에는 Cursor가 293억 달러 밸류에이션으로 시리즈 D 자금 조달에 성공했다. 일일 사용자는 100만 명을 돌파했고, 연간 반복 매출은 10억 달러를 넘겼다. GitHub Copilot은 7월 기준 2,000만 사용자를 확보했고, Fortune 100 기업의 90%가 도입한 것으로 알려졌다. 이는 전년 대비 400% 증가다. 기업 고객은 분기마다 75%씩 늘고 있다.

Stack Overflow의 2025 개발자 설문조사는 현장의 실상을 보여준다. 응답자의 65%가 AI 코딩 도구를 매주 사용하고, 47%는 매일 사용한다. 41%는 자신이 작성하는 코드의 상당 부분이 AI에 의해 생성된다고 답했다. Y Combinator의 2025년 겨울 배치 스타트업 중 25%는 코드베이스의 95%를 AI로 작성했다고 밝혔다. Polaris Market Research는 AI 코딩 어시스턴트 시장이 2024년 49억 1천만 달러에서 2032년 271억 7천만 달러로 성장하며 연평균 23.8%의 성장률을 기록할 것으로 전망했다.

**그림 16. 65%의 개발자들이 주 1회 이상 AI 코딩 도구를 사용 중**




자료: Stack Overflow, 미래에셋증권 리서치센터

숫자는 명확하다. AI 코딩 도구는 틈새에서 주류로 이동했다. 하지만 낙관 뒤에 균열이 보이기 시작했다.

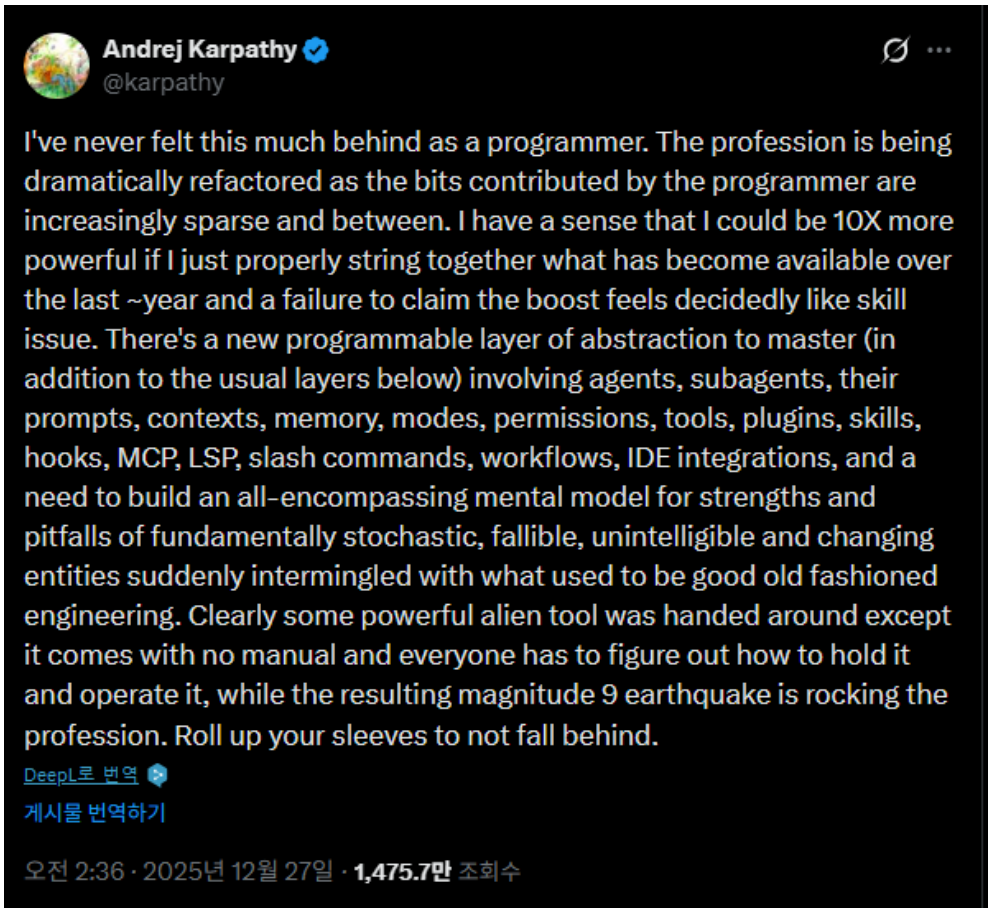
## 2. 변화하는 코딩 패러다임

12월 27일, 안드레이 카파시는 트윗을 올렸다. OpenAI 창립 멤버이자 Tesla의 전 AI 책임자인 그는 "프로그래머로서 이렇게까지 뒤처진 느낌을 받은 적이 없다"고 고백했다. "제대로만 엮어내면 10배 더 강력해질 수 있다는 느낌이 든다. 지난 1년간 사용 가능해진 것들을 제대로 조합하지 못하는 것은 명백히 스킬 이슈다."

그가 말하는 "제대로 엮어내야 할 것"은 새로운 프로그래밍 추상화 계층이다. 에이전트와 서브 에이전트들, 프롬프트와 컨텍스트, 메모리, 모드, 권한, 도구, 플러그인, 훅(hook), MCP, LSP(Language Server Protocol), 슬래시 커맨드, 워크플로우, IDE 통합. 이것들은 모두 기존의 엔지니어링 계층 위에 쌓인 새로운 레이어다. '근본적으로 확률적이고, 오류 가능하고, 이해 불가능하며 변화하는' 모델들을 다루기 위해 필수적인 도구들이다.

카파시는 이를 "확실히 강력하고 새로운 도구가 건네졌지만 매뉴얼이 없다"고 표현했다. 그리고 그것을 어떻게 쥐고 작동시킬지는 모두가 스스로 알아내야 한다. "그 결과 규모 9 지진이 업계를 흔들고 있다. 뒤처지지 않으려면 소매를 걷어붙여라."

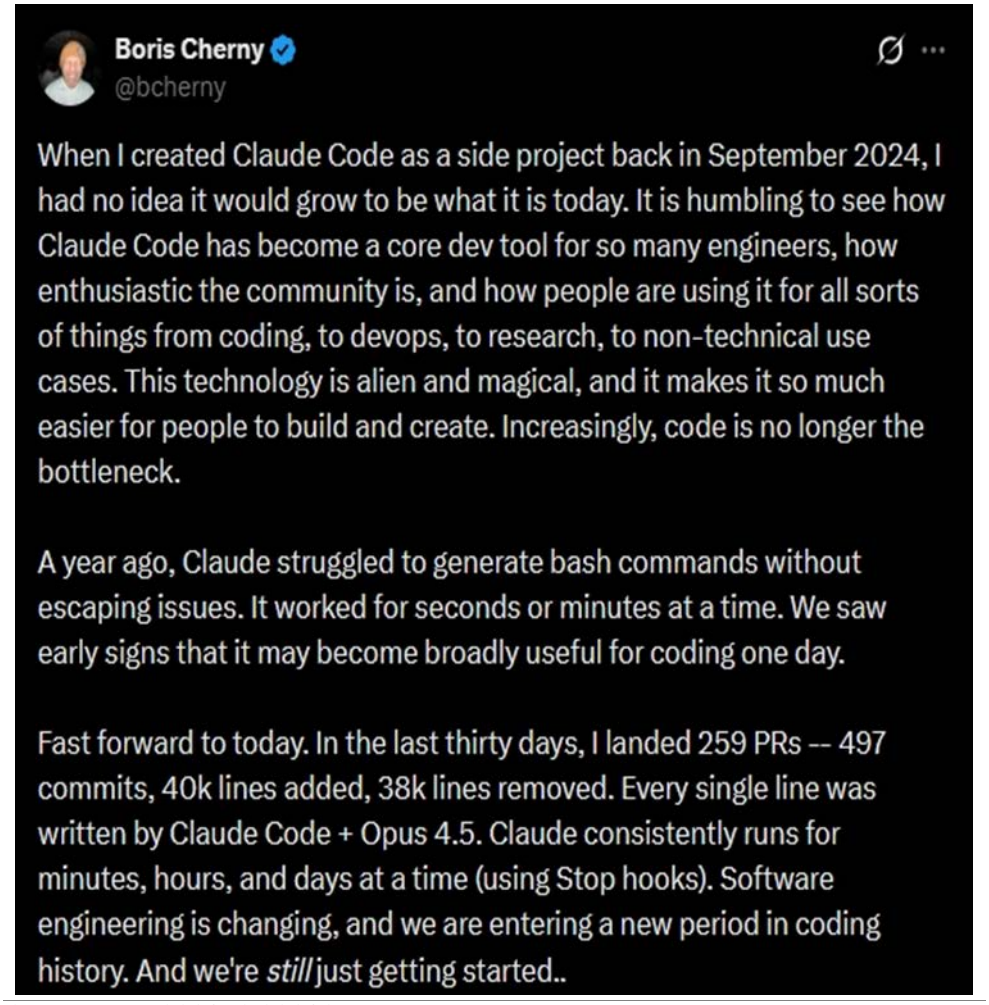
**그림 17. 안드레 카파시가 프로그래머로서 느끼는 AI 코딩의 방대함**




자료: 안드레 카파시, 미래에셋증권 리서치센터

같은 날, Anthropic의 Claude Code 책임자 Boris Cherny도 트윗을 올렸다. 그는 2024년 9월 사이드 프로젝트로 Claude Code를 만들었고, 이제 그것이 "수많은 엔지니어의 핵심 개발 도구"가 됐다고 밝혔다. 그리고 자신의 작업 방식을 공개했다.

**그림 18. 지난 한 달간 모든 코드를 Claude Code와 Opus 4.5가 생성했다는 사실을 공개한 앤트로픽의 Claude Code 개발자**



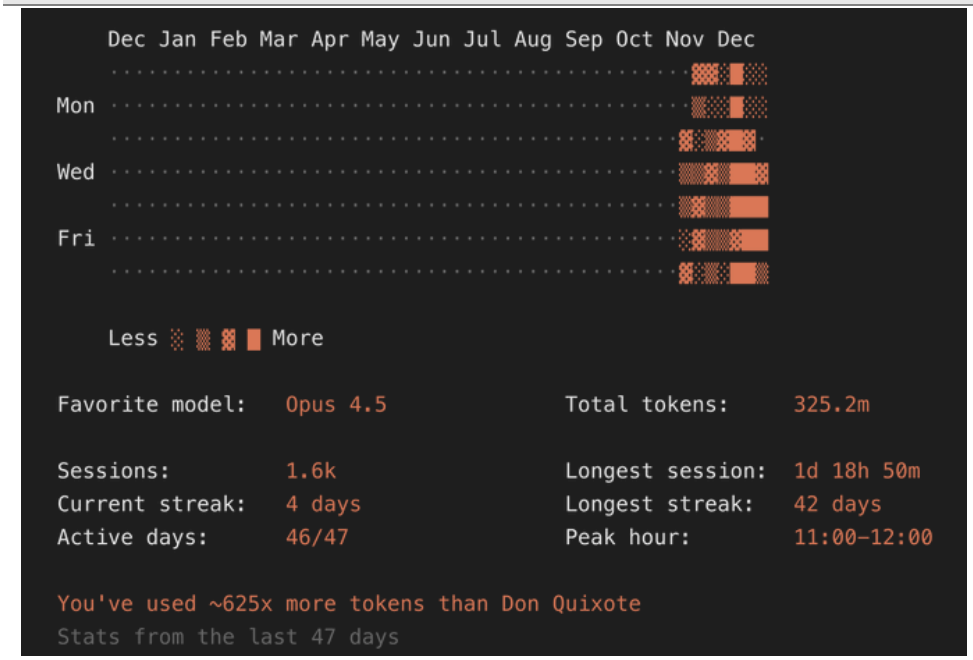

자료: Boris Cherny, 미래에셋증권 리서치센터

"지난 30일간 나는 259개의 PR을 랜딩했다. 497개의 커밋, 4만 라인 추가, 3만 8천 라인 제거. 모든 코드 라인이 Claude Code와 Opus 4.5의 조합으로 작성됐다." 그는 Claude가 이제 몇 분, 몇 시간, 며칠 단위로 일관되게 작동한다고 말했다. "소프트웨어 엔지니어링이 변하고 있다. 우리는 코딩 역사의 새로운 시기로 진입하고 있다."

Cherny는 GitHub 활동 그래프도 함께 공개했다. 커밋 히스토리는 거의 매일 오렌지색 사각형이 빼곡히 채워져 있었고, Opus 4.5 모델을 가장 많이 사용했다. 총 3억 2,520만 개의 토큰과 1,600개의 세션을 사용했고, 가장 긴 세션은 1일 18시간 50분에 달했다. 그의 말처럼, 코드는 더 이상 병목이 아니다.

두 사람의 선언은 같은 현상을 다른 각도에서 포착한다. 카파시는 현업에 있는 개발자들이 배워야 할 새로운 기술 스택의 빠른 분화와 그로 인한 복잡성을 보여줬다면, Cherny는 그 기술을 활용해서 얻을 수 있는 생산성을 보여줬다. 하나는 불안이고, 다른 하나는 증명이다. 같은 날, 업계 최정상의 엔지니어들조차 이 전환의 속도에 압도당하고 있거나, 혹은 그것을 완전히 수용하고 있다.

**그림 19. 그의 작업 내역 요약**

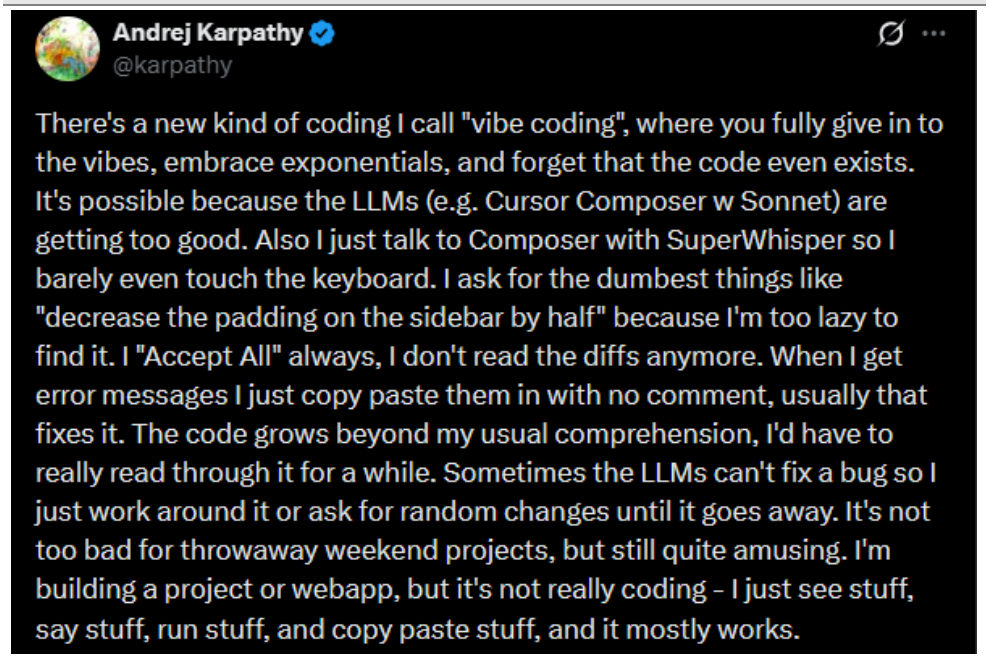



자료: Boris Cherny, 미래에셋증권 리서치센터

바이브 코딩이란 무엇인가? 카파시가 올해 2월 명명한 이 용어는 개발자가 자연어로 원하는 소프트웨어를 설명하면 AI가 코드를 작성하고, 개선하고, 디버깅하는 방식을 가리킨다. diff를 읽지 않고 변경을 수락하고, 에러가 나면 메시지를 복사해 AI에 붙여넣는다. 코드를 "이해"하지 않아도 작동하는 소프트웨어를 만들 수 있다.

그리고 이번 Cherny의 사례는 이것이 더 이상 코딩 아마추어들을 위한 단순한 기능이 아님을 증명한다. 30일간 모든 코드를 AI가 작성했고, 그는 엔지니어에서 프롬프트 엔지니어로 변모했다.

**그림 20. '바이브 코딩'이라는 단어는 안드레 카파시가 처음 만든 것**




자료: 안드레 카파시, 미래에셋증권 리서치센터

## 3. 명과 암: 바이브 코딩에 대한 우려들

### (1) Cursor의 CEO부터 현직 개발자까지

12월 9일, Cursor의 CEO이자 공동창업자 Michael Truell은 Fortune Brainstorm AI 컨퍼런스에서 명확한 선을 그었다. "바이브 코딩은 흔들리는 기반을 만든다." 그는 MIT 출신 4명이 2022년 만든 Cursor를 일일 사용자 100만 명, 연 매출 5억 달러를 달성한 293억 달러 밸류의 기업으로 키웠지만, 동시에 자신의 제품이 가지고 올 수 있는 위험에 대해서도 경고했다.

"바이브 코딩은 아마추어 빌더나 경험 없는 AI 사용자가 후드 아래를 들여다보지 않고 아이디어를 실현하려는 시도를 암시한다. 그들은 날아다니는 맹인이다." Truell은 Cursor가 "양쪽의 최선"이라고 강조했다. 전문가는 코드의 세부사항을 파고들 수 있고, 필요할 때는 AI에 엔드투엔드 작업을 맡길 수도 있다. 하지만 차이는 명확하다. 이해하지 못하는 코드 위에 시스템을 쌓는 것은 위험하다.

같은 주, 개발자 Simon Willison은 블로그에 '당신의 일은 작동한다는 것이 증명된 코드를 전달하는 것'이라는 글을 올렸다. 사실상 동종 업계 사람들에 대한 매서운 비난이다. 그는 AI 보조 소프트웨어 개발 논쟁에서 반복적으로 보이는 현상을 지적했다. "LLM 도구에 힘입은 주니어 엔지니어가 동료나 오픈소스 메인테이너에게 거대하고 테스트되지 않은 PR을 던지고, 코드 리뷰 프로세스가 나머지를 처리해주길 기대한다. 이건 무례하고, 다른 사람의 시간 낭비이며, 솔직히 소프트웨어 개발자로서의 의무 태만이다."

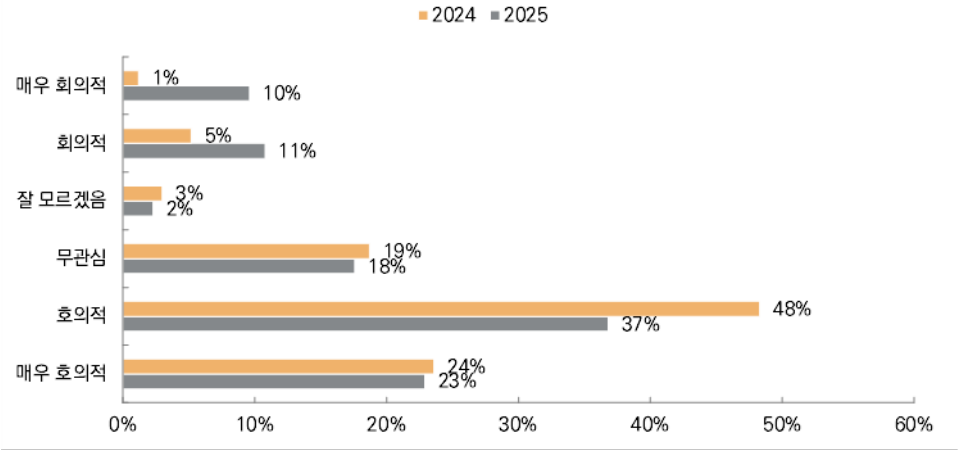
Willison의 요점은 간단하다. 소프트웨어 엔지니어는 코드를 만드는 게 아니라, 작동하는 코드를 전달해야 한다. 그리고 작동한다는 증명도 함께 제공해야 한다. "요즘은 LLM이 코드를 만드는 일을 한다고 주장할 수 있다. (하지만) 우리는 작동하는 코드를 전달해야 한다. 그리고 그것이 작동한다는 증거도 함께." 증명 없이 코드만 던지는 것은 실제 작업을 리뷰어에게 떠넘기는 행위라는 것이다. 같은 개발 직군이라도 세부적인 롤에 따라 AI에 대한 인식이 다르다는 것을 짐작할 수 있다.

### (2) 데이터로도 나타나는 개발자들의 심리 변화

Stack Overflow의 2025 개발자 설문조사는 이런 우려가 과장이 아님을 보여준다. AI 도구에 대한 긍정적 감정은 2023~2024년 70% 이상에서 2025년 60%로 하락했다. 전문 개발자의 61%가 호의적이지만, 학습자는 53%에 불과하다. 더 중요한 수치가 있다. 응답자의 72%가 "바이브 코딩은 내 작업 방식이 아니다"라고 답했다. 5%는 "절대 아니다"라고 강조했다.

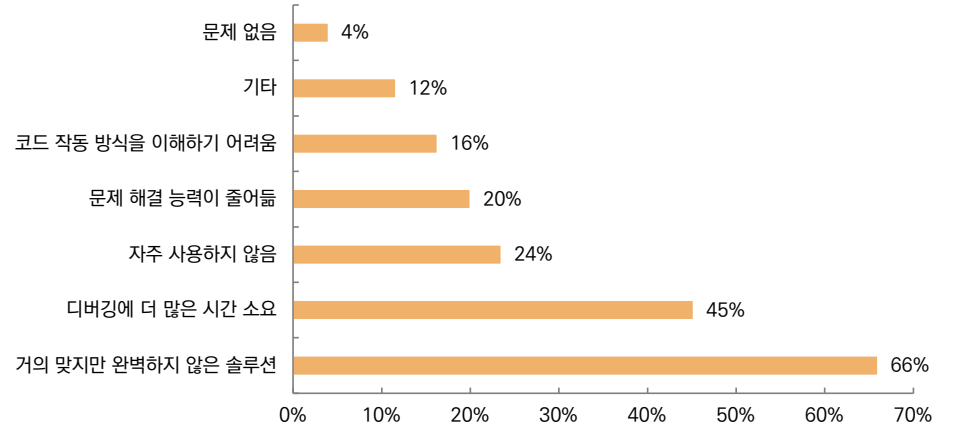
불만의 1순위는 "거의 맞지만 완전히는 아닌" 솔루션이다. 66%가 이것을 꼽았다. AI 코드를 디버깅하는 데 더 많은 시간이 든다는 응답이 45%다. 신뢰 문제는 더 심각하다. 46%가 AI의 정확성을 적극적으로 불신한다. 33%만 신뢰하고, 경험 많은 개발자 중 "고도로 신뢰"하는 비율은 2.6%다. 75%는 AI 답변이 신뢰할 수 없을 때 여전히 사람에게 물어본다고 답했다.

**그림 21. AI 도구에 대한 사람들의 심리 변화. 지난해에 비해 긍정적 인식은 감소, 부정적 인식은 증가한 것을 볼 수 있음**




자료: Stack Overflow, 미래에셋증권 리서치센터

**그림 22. AI 도구에 대해 가지고 있는 불만 1순위는 거의 맞지만 완벽하지는 않은 현재의 솔루션들. 이로 인해 업무가 오히려 과중되는 현상이 발생하고 있음**



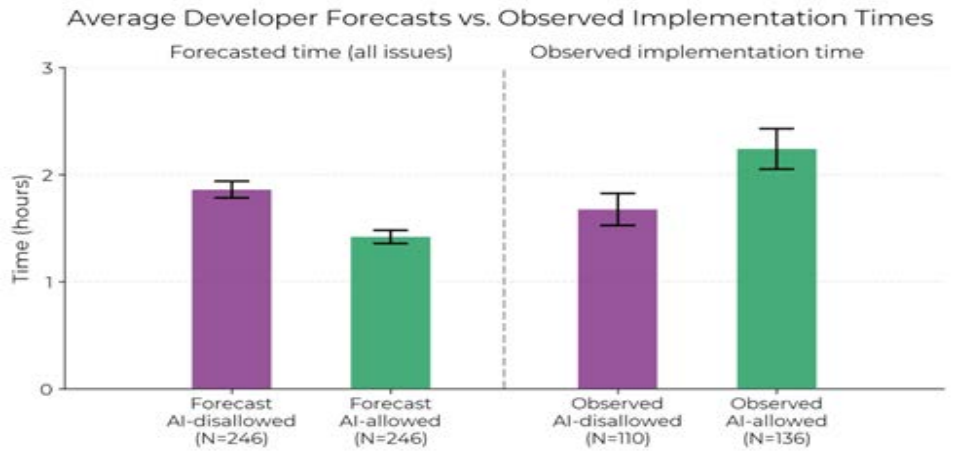

자료: Stack Overflow, 미래에셋증권 리서치센터

고용 시장은 더 가혹한 현실을 보여준다. 8월, Stanford 대학의 Erik Brynjolfsson과 연구팀은 ADP의 급여 데이터를 분석한 논문을 발표했다. 수백만 명의 근로자, 수만 개 기업의 실제 고용 기록을 추적한 결과다. 22세에서 25세 소프트웨어 개발자 고용은 2022년 말 정점 대비 20% 감소했다. AI에 가장 많이 노출된 직군은 같은 기간 13% 상대적 감소를 보였다. 반면 26세 이상 개발자는 6~9% 증가했다. 세대 간 양극화다.

12월, LA Times는 Stanford 졸업생들의 고용 위기를 보도했다. 생체공학과 Jan Liphardt 교수는 "3년 전엔 모든 학부 멘티가 주변 기업에서 좋은 직장을 찾았다. 그게 바뀌었다"고 말했다. Stanford 컴퓨터공학 학위는 취업 황금 티켓에서 동판으로 격하됐다. 한 졸업생은 이렇게 증언했다. "지난 2년간 5학년 석사 지원자가 급증했다. 1년을 더 쓰고, 취업 사이클도 하나 더 돌리는 것이다. 내 친구 절반이 5학년 석사를 하고 있다." 4개월 만에 취업한 어느 졸업생은 현재 AI 도구를 쓰지만 "3명분 일을 하는 느낌"이라고 말했다.

생산성 향상이라는 약속도 흔들린다. 7월, METR는 경험 많은 오픈소스 개발자 16명을 대상으로 무작위 통제 실험을 진행했다. 각자가 기여하는 대형 레포지토리(평균 2만 2천 스타, 100만 라인 이상)에서 246개의 실제 이슈를 선정했다. 평균 2시간 소요 작업이다. 절반은 AI 사용 허용(주로 Cursor Pro + Claude 3.5/3.7 Sonnet), 절반은 불허. 결과는 놀라웠다. AI를 사용한 개발자는 실제로 19% 느려졌다. 하지만 그들 스스로는 20% 빨라졌다고 믿었다.

**그림 23. 개발자들을 대상으로 AI를 사용해 업무를 수행하게 한 결과, AI를 사용했을 때 19% 느려졌지만 개발진들은 더 빨라졌다고 믿음**




자료: METR, 미래에셋증권 리서치센터

12월 Stanford의 Yegor Denisov-Blanch는 대기업 엔지니어링 팀을 추적한 연구를 발표했다. AI 도입 후 코드 품질이 하락했고, 편집이 더 필요했다. 생산성은 거의 변화가 없었다. "코드 라인 수만 측정하면 생산성이 올라간 것처럼 보이지만, 더 깊이 들여다보면 그렇지 않다"고 그는 컨퍼런스에서 말했다.

Coinbase의 경험도 비슷하다. MIT Technology Review 인터뷰에서 플랫폼 책임자 Rob Witoff는 간단한 작업에서 90% 속도 향상을 봤다고 밝혔다. 코드베이스 재구조화, 테스트 작성 같은 일에서는 확실한 성능 개선을 체감한 것이다. 하지만 복잡한 작업의 개선은 미미했다. 더 큰 문제는 다른 데 있었다. "AI 도구는 주니어 개발자가 훨씬 더 많은 코드를 생성하게 한다. 거의 모든 엔지니어링 팀에서 이 코드는 다른 사람, 보통 시니어 개발자가 리뷰해야 한다. 버그를 잡고 품질 기준을 충족시키기 위해서다." 생산성 향상은 그만큼의 리뷰 지옥으로 상쇄된다. 개발 직군 안에서도 상반된 의견이 나타나는 이유다.

이러한 갈등은 게임 산업으로도 번졌다. 12월 18일, Indie Game Awards는 〈Clair Obscur〉의 수상을 취소했다. 올해의 게임과 데뷔 게임 두 부문을 동시 석권한 지 6일 만이었다. 일명 33 원정대라고 불리는 4월 출시된 이 게임은 게임 내 게시판 텍스처가 AI로 생성됐다는 의혹을 받았다. 7월 프로듀서 François Meurisse는 El País 인터뷰에서 "AI를 약간 사용했지만 많지 않다"고 해명했다. 12월 12일 수상 후, 주최 측인 Six One Indie가 사실 확인을 요청했고, Sandfall Interactive는 AI 사용을 인정했다.

회사는 문제의 텍스처가 "플레이스홀더"였고 실수로 최종 빌드에 포함됐으며 이후 패치했다고 설명했다. "AI가 인간 예술을 대체하는 것에 반대한다"는 입장도 밝혔다. 하지만 Six One Indie는 움직이지 않았다. "제출 과정 전반에 걸쳐 생성형 AI에 대한 강경 입장"을 명시한 규정을 위반했다는 것이다. Sandfall 대표들은 초기에 생성형 AI를 사용하지 않았다고 동의했었다.

이 소식에 게임 커뮤니티에도 갑론을박이 벌어졌다. 지지자들은 "AI는 언리얼 엔진처럼 그냥 도구일 뿐"이라고 주장하는 한편, 비판자들은 "플레이스홀더였다는 변명"과 초기 부인을 문제 삼았다. 질문은 명확하다. 이 경우에는 동의서를 작성했고, 이로 인한 규정 위반이라는 점이 더 중요했지만, AI를 "보조 도구"로 사용하는 것과 "창작의 핵심"으로 의존하는 것의 경계는 과연 어디인지, 그리고 그 경계를 누가 정의할 수 있는지에 대한 질문도 동시에 던지는 사건이었다.

Clair Obscur 논란은 결국 앞선 예시들과 같은 구조다. Cursor CEO가 경고한 "흔들리는 기반", Willison이 지적한 "증명 없는 코드", Stanford 졸업생들이 직면한 "황금에서 청동으로의 전락" 등, AI 도구가 확산될수록 투명성과 품질, 윤리의 경계를 둘러싼 산업 전반의 마찰은 심화되고 있으며, 사실상 개발자들이 속한 모든 산업에서 언제든 일어날 수 있다. 숫자는 확장을 증명하지만, 현장은 균열을 목격하고 있다.

## 4. 매뉴얼이 없는 도구는 장인의 손을 탄다

안드레 카파시의 "규모 9 지진" 발언은 과장이 아니다. 2025년은 프로그래밍 직업이 근본적으로 재정의된 해로 기록될 것이다. 바이브 코딩은 더 이상 실험이 아니라 산업 표준이 되어가고 있다. 41%의 코드가 AI에 의해 생성되고, Y Combinator 스타트업의 25%가 코드베이스의 95%를 AI로 채우는 현실은 불가역적 전환이다.

하지만 전환은 마찰 없이 오지 않는다. Cursor CEO는 "흔들리는 기반"을 경고했고, Simon Willison은 "증명 없는 코드"의 위험을 지적했다. Stanford 졸업생들은 취업의 문이 실시간으로 닫혀가는 것을 목격하고 있다. 22세에서 25세 개발자의 고용은 20% 감소했고, 경험 많은 개발자조차 AI를 사용하면 주관적으로는 20% 빨라졌다고 믿지만 실제로는 19% 느려졌다. Coinbase는 주니어의 코드 폭증이 시니어의 리뷰 지옥을 만든다는 것을 배웠다. Clair Obscur는 AI 사용 자체에 대한 업계의 반발 심리가 어느 정도인지를 보여줬다.

낙관과 회의가 공존한다. Boris Cherny는 30일간 259개의 PR을 AI로 완성하며 "코드는 더 이상 병목이 아니다"라고 선언했지만, 동시에 72%의 개발자는 '바이브 코딩은 내 방식이 아니'라고 생각하고 있으며, 46%는 AI를 불신한다. 66%는 거의 맞지만 완전하지는 않은 솔루션에 불만을 느끼고 있고, 개발자 커뮤니티의 긍정 비율은 70%에서 60%로 하락했다.

카파시는 뒤처지지 않으려면 매뉴얼 없는 외계 도구를 최대한 활용해야 하며, 이를 위해 소매를 걷어붙여야 한다고 촉구했다. 하지만 어디로 가야 할지는 아무도 모른다. 새로운 추상화 계층을 배워야 하는지, 에이전트나 MCP, 워크플로우를 마스터해야 하는지, 아니면 Cursor CEO의 경고처럼 이해하지 못하는 코드 위에 시스템을 쌓는 것을 경계하기 위해 직접 코딩하는 능력을 길러야 할지, 그 누구도 갈피를 잡지 못하고 있다.

확실한 것은 전환이 진행 중이라는 사실뿐이다. 바이브 코딩은 확장하고 있고, 동시에 마찰을 일으키고 있다. 어떤 이는 10배 생산성을 경험하고, 어떤 이는 일자리를 잃는다. 고전적인 코딩 능력과 바이브 코딩이 열어준 새로운 지평에서 자신에게 필요한 능력을 잘 선택해서 조합해 나가야 하며, 그 선택 과정과 결과는 각 개인의 손에 달려 있다. 코딩의 새로운 지평을 여는 과정에서,

# IV. AI Peer Table

**표 4. 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 | 273.76 | 5,805 | 1.0 | -1.8 | 7.7 | 36.4 | 7.6 | 9.8 | 01-30 | 엔드유저 데이터 보유 및 서비스 배포능력 |
| 테슬라 | TSLA US | 459.64 | 2,194 | -6.0 | 6.9 | 3.7 | 42.0 | 6.5 | 13.8 | 01-29 | FSD v13 공공 출시 및 1Q25 중국 출시 가능성 |
| Pony AI | PONY US | 15 | 9 | -8.4 | 9.1 | -34.3 | 9.2 | 8.7 | 4.0 | 03-25 | 중국의 자율주행 선도기업 중 하나 |
| 삼성전자 | 005930 KS | 120,600 | 714 | 8.2 | 18.8 | 43.7 | 102.6 | 130.5 | 130.5 | 01-07 | 온디바이스 AI, HBM, 파운드리 영역에서의 가능성 |
| 듀오링고 | DUOL US | 180.17 | 12 | -2.8 | -5.9 | -45.2 | -56.2 | -46.0 | -44.4 | 02-27 | AI 캐릭터와 상호작용하는 학습 플랫폼 |
| 크래프톤 | 259960 KS | 245,500 | 12 | -2.0 | -4.5 | -16.1 | -32.4 | -21.4 | -21.4 | 02-11 | PUBG, inZOI 등 자사 게임에 AI CPC 기능 탑재 |
| 쇼피파이 | SHOP US | 167.88 | 314 | -1.1 | 5.8 | 12.7 | 47.7 | 55.2 | 57.9 | 02-11 | 이커머스 콘텐츠 생성의 자동화(Shopify Magic AI) |
| 코인베이스 | COIN US | 233.77 | 90 | -5.7 | -14.3 | -30.0 | -33.9 | -12.0 | -5.9 | 02-13 | AI 에이전트간 자율 암호화폐 거래 인프라 구축 |
| **비즈니스 효율화** | | | | | | | | | | | |
| 크라우드스트라이크 | CRWD US | 475.91 | 172 | -1.5 | -6.5 | -2.6 | -4.7 | 34.1 | 39.1 | 03-04 | 기업 고객 대상 엔드포인트 보안(Charlotte AI) |
| IBM | IBM US | 305.74 | 410 | 1.0 | -0.9 | 9.9 | 6.8 | 40.8 | 42.7 | 01-28 | 기업 고객 대상 AI 모델 개발, 배포 플랫폼(watsonx) |
| 액센츄어 | ACN US | 271.34 | 242 | 0.2 | 8.5 | 10.6 | -7.1 | -22.2 | -21.2 | 03-19 | 각 산업별 맞춤형 AI 에이전트를 만들기 위한 첨병 |
| SAP | SAP GY | 208.45 | 432 | -0.0 | -0.0 | -7.6 | -19.2 | -12.2 | -11.0 | 01-29 | 기업용 SW 솔루션에 AI를 내장한 에이전트(Joule) |
| 인포시스 | INFO IN | 1,644.70 | 109 | -2.7 | 5.4 | 15.8 | 3.8 | -11.7 | -9.9 | 01-14 | 각 산업별 맞춤형 AI 에이전트를 만들기 위한 첨병 |
| 일루미나 | ILMN US | 134.48 | 29 | -2.3 | 2.3 | 46.4 | 41.9 | -0.6 | 0.6 | 02-06 | AI 기반 첨단 유전체 분석 |
| HD현대 | 267250 KS | 186,700 | 15 | -3.5 | -5.2 | 20.1 | 45.0 | 144.6 | 144.6 | 02-06 | AI, 로보틱스, 디지털 트랜스포메이션 관련 노력 |
| **AI Ops** | | | | | | | | | | | |
| 클라우드플레어 | NET US | 200.70 | 101 | -0.7 | 0.2 | -8.2 | 3.4 | 81.4 | 86.4 | 02-06 | 실시간 AI 추론에 CDN 수요증가 가능성(Workers AI) |
| 팔란티어 | PLTR US | 184.18 | 630 | -5.1 | 9.3 | 3.0 | 40.9 | 132.9 | 143.5 | 02-03 | 비즈니스 현장의 의사결정을 돕는 AI 플랫폼(AIP) |
| **파운데이션 모델** | | | | | | | | | | | |
| 알파벳 | GOOGL US | 313.56 | 5,437 | 1.2 | -2.0 | 28.6 | 75.9 | 63.3 | 66.3 | 02-04 | OpenAI와의 AGI 경쟁이 가능한 업체(Gemini) |
| 메타 플랫폼스 | META US | 658.69 | 2,383 | -0.4 | 1.7 | -11.3 | -10.1 | 10.2 | 12.8 | 01-29 | 오픈소스 AI 개발의 선구자(Llama 등) |
| 알리바바 | 9988 HK | 144.40 | 509 | -1.9 | -4.7 | -18.4 | 31.5 | 79.0 | 78.1 | 02-20 | 중국의 CSP이자 중국의 대표적인 언어모델(Qwen) |
| 네이버 | 035420 KS | 243,000 | 38 | 2.7 | -1.2 | -9.5 | -7.4 | 22.8 | 22.8 | 02-06 | 한국형 LLM 개발 및 AI 서비스(HyperCLOVA) |

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

### 표 5. 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 | 487.10 | 5,195 | 0.4 | -1.0 | -5.2 | -1.4 | 14.0 | 16.4 | 01-29 | OpenAI 모델 라이선스 독점권 보유한 세계 2위 CSP |
| 아마존 닷컴 | AMZN US | 232.07 | 3,560 | 1.6 | -0.5 | 4.5 | 3.9 | 3.7 | 5.8 | 02-06 | 최고의 CSP로서 Anthropic에 수 조원 투자 |
| 오라클 | ORCL US | 195.38 | 806 | -1.5 | -3.3 | -30.8 | -6.7 | 16.8 | 18.4 | 03-10 | AI 데이터센터 capex 경쟁에 진입한 전통적 강자 |
| 소프트뱅크 | 9984 JP | 4,412.00 | 232 | -0.5 | 4.9 | -5.5 | 68.0 | 92.9 | 92.9 | 02-12 | '비전펀드'는 AI에 집중. 일본 최고의 AI 슈퍼컴퓨터. |
| **하드웨어 인프라** | | | | | | | | | | | |
| 엔비디아 | NVDA US | 188.22 | 6,564 | 2.5 | 6.3 | 3.5 | 19.3 | 37.4 | 40.2 | 02-26 | AI 모델 훈련 및 추론에 필수인 GPU계 현존 최강자 |
| 브로드컴 | AVGO US | 349.39 | 2,377 | 2.3 | -13.1 | 6.8 | 30.2 | 45.8 | 52.1 | 03-06 | 이더넷 기반 네트워킹 반도체의 최강자 |
| AMD | AMD US | 215.61 | 504 | 0.3 | -0.9 | 33.6 | 49.9 | 72.2 | 78.5 | 02-04 | 서버용 CPU의 최강자. AI 가속기 분야 패스트팔로어 |
| 인텔 | INTC US | 36.68 | 262 | 0.9 | -9.6 | 6.4 | 61.7 | 80.7 | 82.9 | 01-30 | 파운드리로서 소버린 AI 미 지정학적 가치 부상 |
| 마이크론 | MU US | 294.37 | 475 | 6.5 | 24.5 | 79.8 | 136.4 | 233.7 | 250.9 | 03-20 | SK하이닉스 추격 중인 "미국"의 메모리 반도체 |
| SK하이닉스 | 000660 KS | 642,000 | 475 | 11.6 | 21.4 | 87.8 | 123.8 | 278.8 | 278.8 | 01-22 | HBM 부문 전세계 최강자 |
| 시놉시스 | SNPS US | 478.97 | 132 | -0.5 | 14.6 | -0.5 | -4.7 | -2.4 | -1.3 | 02-26 | 반도체 EDA 부문 리딩 업체, AI 에이전트 적극 활용 |
| 케이던스 디자인 시스템 | CDNS US | 317.71 | 124 | 0.0 | 1.9 | -8.8 | 4.1 | 4.2 | 5.7 | 02-18 | 반도체 EDA 부문 리딩 업체 |
| 암페놀 | APH US | 136.90 | 240 | 1.3 | -2.7 | 13.4 | 40.6 | 95.0 | 98.6 | 01-28 | 엔비디아향 AI용 서버 내 고속 I/O 및 커넥터 제조 |
| 버티브 | VRT US | 165.62 | 91 | -0.4 | -7.9 | 15.6 | 30.3 | 44.1 | 46.0 | 02-12 | 액체 냉각 방식에 있어 글로벌 선두업체 |
| Arm | ARM US | 110.51 | 168 | -2.5 | -18.5 | -21.0 | -33.2 | -14.5 | -10.4 | 02-05 | AI 가속기 설계 위한 다수의 IP를 소유한 팹리스 |
| 마벨 테크놀로지스 | MRVL US | 85.76 | 104 | 1.1 | -4.1 | 4.2 | 11.3 | -24.3 | -22.1 | 03-05 | 데이터 인프라용 네트워킹 및 스토리지 ASIC 설계 |
| 코어위브 | CRWV US | 74.92 | 54 | -11.7 | 2.5 | -38.9 | -53.2 | - | - | 02-13 | AI용 클라우드 인프라 제공업체 |
| 중과서광 | 603019 CH | 86.50 | 26 | -1.3 | -12.8 | -27.5 | 22.9 | 17.3 | 20.1 | 04-23 | 중국과학원과 긴밀 관계를 유지 중인 HPC 제조업체 |
| SMIC | 981 HK | 71.25 | 128 | 3.4 | 3.6 | -10.4 | 59.4 | 118.6 | 124.1 | 02-11 | 7나노 공정을 달성한 중국 유일의 미세공정 파운드리 |
| TSMC | 2330 TT | 1,530.00 | 1,801 | 2.0 | 5.9 | 16.9 | 44.4 | 41.7 | 43.7 | 01-15 | 명실상부 반도체 파운드리 최강자 |
| 폭스콘 | 2317 TT | 229.50 | 146 | 2.5 | 1.8 | 6.3 | 47.7 | 28.5 | 29.2 | 03-16 | GPU 기반 보드 및 서버 시스템 조립 담당, 대만 업체 |
| 위스트론 | 3231 TT | 145.00 | 21 | 0.0 | 0.3 | 3.2 | 18.4 | 44.3 | 44.3 | 01-16 | GPU 기반 보드 및 서버 시스템 조립 담당, 대만 업체 |
| Alchip | 3661 TT | 3,510.00 | 13 | 9.0 | 6.2 | 1.3 | 14.5 | 10.2 | 8.0 | 03-02 | 빅테크들의 AI 가속기 설계를 돕는 대만의 팹리스 |
| **에너지 인프라** | | | | | | | | | | | |
| GE 베르노바 | GEV US | 663.46 | 258 | 0.2 | 10.6 | 10.2 | 27.8 | 99.1 | 102.0 | 01-22 | AI 데이터센터용 가스터빈 인프라 솔루션 |
| 슈나이더 일렉트릭 | SU FP | 236.30 | 230 | -0.1 | 2.3 | 0.9 | 2.7 | -0.2 | -0.2 | 02-26 | AI 데이터센터 전력관리 및 냉각시스템 통합 솔루션 |
| 비스트라 | VST US | 161.84 | 79 | 0.2 | -9.4 | -18.1 | -16.8 | 16.5 | 18.0 | 02-27 | 미국 최대 독립 전력 생산 사업자 |
| 블룸 에너지 | BE US | 88.41 | 30 | -4.2 | -19.1 | 20.1 | 298.6 | 286.1 | 298.1 | 02-27 | AI 데이터센터용 연료전지 솔루션 |
| 넥스트에라 에너지 | NEE US | 80.27 | 240 | 0.3 | -7.0 | 6.0 | 14.9 | 14.8 | 15.5 | 01-23 | AI 데이터센터용 엔드투엔드 전력 파트너 |

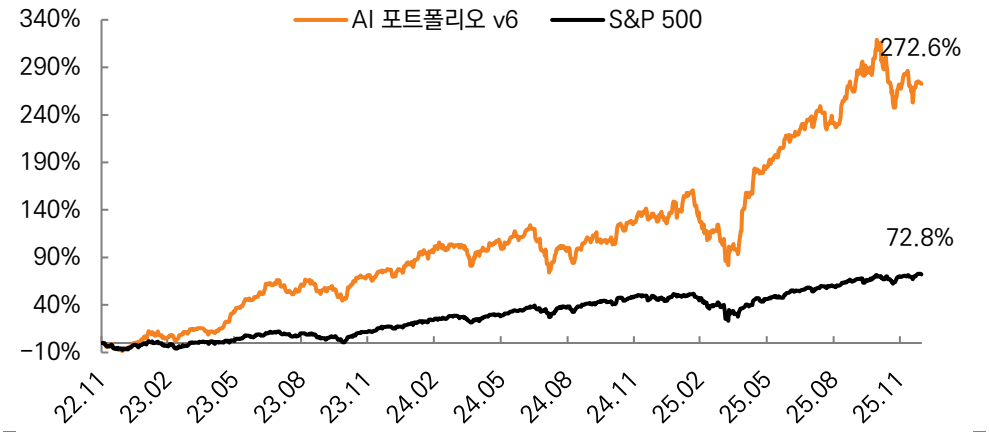
자료: Bloomberg, 미래에셋증권 리서치센터
주: 실적발표 일정은 변경될 수 있음.

# V. Appendix: Charts

**그림 24. ChatGPT 출시일 이후, S&P 500에서 Magnificent Seven이 차지하는 비중(시가총액 기준)**

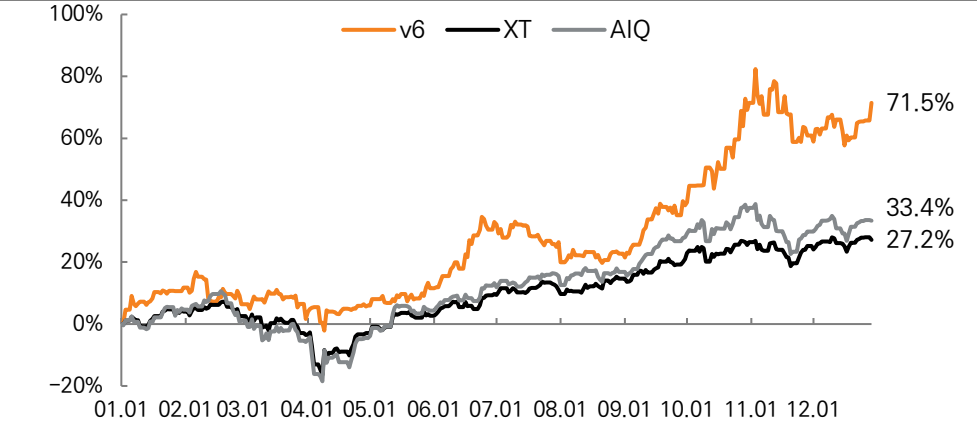
자료: Bloomberg, 미래에셋증권 리서치센터

**그림 25. ChatGPT 출시일 이후, AI 밸류체인에 속한 유니버스 종목들의 평균 수익률(동일가중 방식)**




자료: Bloomberg, 미래에셋증권 리서치센터

**그림 26. 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 6. 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(New) | Gemini 3 Flash Preview | 61.1 | 구글 | 독점 폐쇄소스 | 2025년 1월 |
| 6(▼1) | Grok 4 | 60.5 | xAI | 독점 폐쇄소스 | 2024년 12월 |
| 7(▼1) | Claude 4.1 Opus | 60.0 | 앤트로픽 | 독점 폐쇄소스 | 2025년 3월 |
| 8(▼1) | Claude 4 Opus | 58.8 | 앤트로픽 | 독점 폐쇄소스 | 2025년 3월 |
| 9(▼1) | GPT-5.2 Pro(xhigh) | 57.4 | OpenAI | 독점 폐쇄소스 | |
| 9(▼1) | 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 7. 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 | 1490 | 구글 | 독점 폐쇄소스 | 2025년 1월 |
| 2(New) | gemini-3-flash | 1478 | 구글 | 독점 폐쇄소스 | 2025년 1월 |
| 3(▲1) | grok-4.1-thinking | 1477 | xAI | 독점 폐쇄소스 | - |
| 4(New) | claude-opus-4-5-20251101-thinking-32k | 1469 | 앤트로픽 | 독점 폐쇄소스 | 2025년 8월 |
| 5(▲2) | claude-opus-4-5-20251101 | 1467 | 앤트로픽 | 독점 폐쇄소스 | 2025년 8월 |
| 6(▲1) | grok-4.1 | 1464 | xAI | 독점 폐쇄소스 | - |
| 7(New) | gemini-3-flash (thinking-minimal) | 1463 | 구글 | 독점 폐쇄소스 | 2025년 1월 |
| 8(-) | gpt-5.1-high | 1455 | OpenAI | 독점 폐쇄소스 | 2024년 9월 |
| 9(New) | gemini-2.5-pro | 1451 | 구글 | 독점 폐쇄소스 | 2025년 1월 |
| 10(▼5) | claude-sonnet-4-5-20250929-thinking-32k | 1450 | 앤트로픽 | 독점 폐쇄소스 | 2025년 3월 |

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

**표 8. LLM 벤치마크 LiveBench의 성능 상위 모델 10개(동일 모델 생략)**

| 순위 (전 주 변동) | 모델명 | 전체 평균 점수 | 기관명 | 라이선스 | 지식 컷오프 |
|---|---|---|---|---|---|
| 1(▲1) | Claude 4.5 Opus Thinking High Effort | 76.20 | 앤트로픽 | 독점 폐쇄소스 | 2025년 7월 |
| 2(▼1) | GPT-5.1 Codex Max | 75.63 | OpenAI | 독점 폐쇄소스 | 2024년 9월 |
| 3(▲1) | Gemini 3 Pro Preview High | 75.22 | 구글 | 독점 폐쇄소스 | 2025년 1월 |
| 4(▲1) | GPT-5.2 High | 74.12 | OpenAI | 독점 폐쇄소스 | |
| 5(▲1) | GPT-5 Pro | 73.82 | OpenAI | 독점 폐쇄소스 | 2024년 9월 |
| 6(New) | Gemini 3 Flash Preview High | 73.74 | 구글 | 독점 폐쇄소스 | 2025년 1월 |
| 7(New) | GPT-5.1 High | 71.41 | OpenAI | 독점 폐쇄소스 | 2024년 9월 |
| 8(-) | GPT-5.1 Codex | 69.51 | OpenAI | 독점 폐쇄소스 | 2024년 9월 |
| 9(New) | Claude 4.1 Opus Thinking | 67.22 | 앤트로픽 | 독점 폐쇄소스 | 2025년 3월 |
| 10(New) | DeepSeek V3.2 Thinking | 66.22 | DeepSeek | | |

자료: LiveBench, Huggingface, 미래에셋증권 리서치센터

### 표 9. 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(New) | GPT-5.2 (2025-12-11) (high reasoning) | 71.80 | OpenAI | 2025-12-11 |
| 4(▼1) | Claude 4.5 Sonnet (20250929) | 70.60 | 앤트로픽 | 2025-09-29 |
| 5(New) | GPT-5.2 (2025-12-11) | 69.00 | OpenAI | 2025-12-11 |
| 6(▼2) | Claude 4 Opus (20250514) | 67.60 | 앤트로픽 | 2025-08-02 |
| 7(New) | GPT-5.1-codex (medium reasoning) | 66.00 | OpenAI | 2025-11-24 |
| 7(New) | GPT-5.1 (2025-11-13) (medium reasoning) | 66.00 | OpenAI | 2025-11-20 |
| 9(▼4) | GPT-5 (2025-08-07) (medium reasoning) | 65.00 | OpenAI | 2025-08-07 |
| 10(▼4) | Claude 4 Sonnet (20250514) | 64.93 | 앤트로픽 | 2025-05-21 |

자료: SWE-Bench, 미래에셋증권 리서치센터

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

| 순위 (전주 변동) | 모델 | 성능(%) | 비용($) | 평가 기준일 | 실험 환경 구분 |
|---|---|---|---|---|---|
| 1(-) | Claude Opus 4 High | 66 | 313.83 | 2025년 5월 | TAU-bench Few Shot |
| 2(-) | Claude Opus 4.1 High | 62 | 298.58 | 2025년 8월 | TAU-bench Few Shot |
| 3(-) | o4-mini-high | 60 | 18.92 | 2025년 4월 | TAU-bench Few Shot |
| 4(-) | Claude-3.7 Sonnet High | 60 | 37.23 | 2025년 2월 | TAU-bench Few Shot |
| 5(-) | o4-mini-high | 56 | 11.36 | 2025년 4월 | TAU-bench Tool Calling |
| 6(-) | Claude-3.7 Sonnet | 56 | 42.11 | 2025년 2월 | HAL Generalist Agent |
| 7(-) | GPT-4.1 | 56 | 42.58 | 2025년 4월 | TAU-bench Few Shot |
| 8(-) | Claude Opus 4 | 56 | 363.30 | 2025년 5월 | TAU-bench Few Shot |
| 9(-) | o3 Medium | 54 | 14.56 | 2025년 4월 | TAU-bench Tool Calling |
| 10(-) | Claude Opus 4.1 | 54 | 180.49 | 2025년 8월 | HAL Generalist Agent |

자료: HAL, 미래에셋증권 리서치센터
주: TAU-bench Airline은 AI 에이전트를 통한 개인 업무 자동화 능력을 평가하는 벤치마크

## 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 엔비디아 60주 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.