Market Insight | 2025.11.28
그림: Midjourney(프로처럼 그림을 그리는 바나나들의 그림 공장)

MIRAE ASSET 미래에셋증권

AI Bi-weekly

구글의 풀스택과 TPU 독립

한종목
chongmok.han@miraeasset.com

김은지
eunji.kim.a@miraeasset.com

CONTENTS

Highlight of the Week 3
구글의 풀스택 역습과 AI 인프라의 2차전 3
30분의 한계를 넘어, 기억하는 에이전트로 4

I. 반복되고 강조되는 구글의 풀스택은 엔비디아를 불안하게 해요? 5
1. Gemini 3.0 Pro... Power Overwhelming 5
2. 나노 바나나 프로(NBP): 장난감에서 산업적 도구로 11
3. 풀스택의 힘: 인프라의 승리 16
4. TPU vs GPU 22
5. 승자의 조건: 결국에는 '메모리'와 '광통신'으로 흐른다 26

II. AI Agents 31
1. 쇼핑 에이전트와 30분의 현실 31
2. 기억의 부재와 만족의 상대성 33
3. 처방 1: 조직적 기억, 개인적 의도 35
4. 처방 2: 세상을 보는 눈과 뇌 37
5. 결론: 기억을 선점하는 자가 시대를 지배한다 38

III. AI Peer Table 40

IV. Appendix: Charts 42

Highlight of the Week

구글의 풀스택 역습과 AI 인프라의 2차전

2025년 11월, 구글의 Gemini 3.0 Pro 출시는 단순 모델 업데이트가 아닌 AI 산업의 '계단식 도약'을 알리는 신호탄이었습니다. Artificial Analysis 인덱스 1위 탈환, HLE 벤치마크 37.5% 달성, 그리고 GPQA 92% 기록은 "데이터 고갈로 인한 AI 성능 정체" 우려를 불식시키며 스케일링 법칙이 여전히 유효함을 증명했습니다. 특히 VPCT 91% 기록으로 입증된 '공간 추론 능력'은 AI가 텍스트를 넘어 물리적 세계를 이해하기 시작했음을 시사하며, 로보틱스 산업의 기술적 병목을 해소할 열쇠가 될 것으로 사료됩니다. 또한, 자판기 운영 시뮬레이션에서 보여준 전략적 소싱과 동맹, 배신 행위는 AI가 자율적 경제 주체로서 실물 경제에 개입할 준비를 마쳤음을 방증합니다.

Gemini 3.0의 성공 이면에는 구글만의 독보적인 '풀스택' 인프라가 존재합니다. 10조 파라미터급 거대 모델을 학습시키기 위해 구글은 엔비디아 GPU가 아닌 자체 TPU v7 (Ironwood) 100%로 훈련을 수행했습니다. 9,216개의 칩을 단일 슈퍼포드로 묶어 1.77PB의 메모리를 공유하는 이 아키텍처는, JAX와 Pathways, 그리고 XLA라는 소프트웨어 해자와 결합되어 엔비디아 대비 1/3 비용으로 동일한 컴퓨팅 파워를 구현했습니다. 이는 향후 AI 가격 전쟁에서 구글이 생존할 수 있는 압도적 기초 체력을 제공합니다.

데이터센터의 전력이 고정된 상황에서 칩의 가치는 '와트당 토큰 생산량'으로 결정됩니다. 효율이 떨어지는 자체 칩 사용은 비용 절감이 아닌 매출 기회비용의 손실을 의미하므로, 시장은 엔비디아와 구글 등 소수의 '슈퍼 칩' 위주로 재편될 것입니다.

TPU의 약진이 엔비디아의 붕괴를 의미하지는 않습니다. 엔비디아는 범용성과 최첨단 추론 시장을, 구글은 거대 모델 학습과 내부 워크로드를 전담하는 '명확한 시장 분할'이 이루어질 전망입니다. 구글의 TPU 외부 세일즈와 웨이퍼 선점 전략은 엔비디아와의 협상 레버리지를 극대화하는 카드로 활용될 것입니다.

투자자 입장에서 구글과 엔비디아 중 누가 승리할지 베팅하는 것은 리스크가 큽니다. 오히려 확실한 알파는 이 거대한 연산 괴물들이 작동하기 위해 필수적인 '메모리'와 '광통신'에 있습니다. AI가 '월드 모델'로 진화할수록 필요한 메모리 용량은 지수함수적으로 증가합니다. HBM은 단순 부품이 아닌 시스템 내 가장 비싼 부동산이 되었으며, 이는 메모리 업체들의 슈퍼사이클을 예고합니다. 또한 구리선의 물리적 한계로 인해 데이터센터 연결은 '반드시 광학'의 단계로 진입했습니다. 1.6T, 3.2T 시대로의 전환은 관련 인프라 기업들에게 구조적인 수혜를 안겨줄 것입니다.

결론적으로, 구글의 풀스택 역습은 엔비디아 독주 체제에 균열을 내는 동시에, AI 인프라 시장 전체의 파이를 폭발적으로 키우는 촉매제가 되고 있습니다. 우리는 지금 칩 중심의 1차전을 지나, 대역폭과 효율성이 지배하는 2차전의 초입에 서 있습니다.

30분의 한계를 넘어, 기억하는 에이전트로

구글과 앤트로픽 등 경쟁사들이 수직적 전문성을 앞세워 시장을 잠식해 들어오는 가운데, OpenAI는 '쇼핑 리서치' 기능을 선보이며 B2C 에이전트 영역으로의 전면적인 확장을 선언했습니다. 이는 단순한 챗봇을 넘어 소비자의 일상과 지갑을 여는 에이전트 시대로의 진입을 의미하지만, 동시에 소비자의 취향과 전문성을 고려한 깊이 있는 니즈를 충족시키기에는 여전히 '기억'과 '맥락'이 부족하다는 한계 또한 명확히 드러냈습니다.

현재의 AI 에이전트는 30분 내외의 단기 과제에서는 인간을 능가하지만, 그 이상의 호흡이 필요한 복합 과제에서는 급격히 무력해지는 '30분의 벽'에 갇혀 있습니다. 그 원인은 인간의 해마와 같은 장기 기억 메커니즘의 부재와 텍스트 학습의 한계인 물리적 상식의 결핍에 있으므로, 파편화된 대화 로그가 아닌 사용자의 취향과 업무 맥락을 구조화하여 저장하는 시스템 없이는 에이전트가 진정한 비서로 거듭날 수 없습니다.

다행히 빅테크들은 이 문제를 해결하기 위한 구체적인 로드맵을 제시하고 있습니다. 기억 측면에서는 OpenAI가 쇼핑 에이전트를 통해 개인의 의도를, 마이크로소프트는 'Work IQ'를 통해 기업의 업무 맥락을 구조화된 기억으로 구축하고 있으며, 감각 측면에서는 Meta의 'SAM 3'가 3D 객체를 인식하는 눈을, 딥마인드의 'SIMA 2'가 물리 법칙을 체득하는 뇌를 만들고 있습니다. 결국 다가올 4,000조 원 규모의 에이전트 시장 패권은 단순한 연산 능력이 아니라, 사용자의 맥락을 기억하고 물리적 세계를 이해하는 아키텍처를 선점하는 기업이 쥐게 될 것입니다.

표 1. AI 관련 주요 일정

일	월	화	수	목	금	토
23	24	25	26	27	28	29
		· PONY 실적 · BABA 실적				
30	1	2	3	4	5	6
		· MRVL 실적 · CRWD 실적				
7	8	9	10	11	12	13
		· ORCL 실적(예)	· SNPS 실적			

자료: Bloomberg, 미래에셋증권 리서치센터

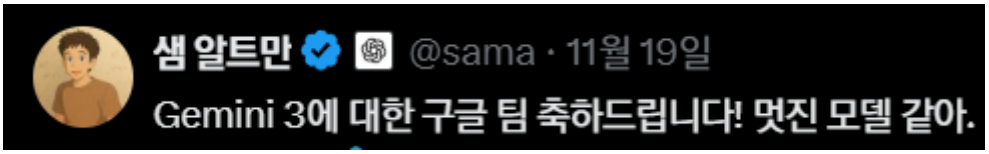
I. 반복되고 강조되는 구글의 풀스택은 엔비디아를 불안하게 해요?

1. Gemini 3.0 Pro... Power Overwhelming

(1) 11월의 특이점: 계단식 도약

2025년 11월은 AI 역사상 가장 밀도 높은 업그레이드 주기로 기록될 전망이다. **12일 OpenAI의 GPT-5.1을 시작으로 17일 xAI의 Grok 4.1, 18일 구글 Gemini 3 Pro, 그리고 24일 Anthropic의 Opus 4.5가 연이어 쏟아져 나왔다**. 이 숨 가쁜 릴레이의 정점은 단연 구글의 Gemini 3 Pro Preview였다. 이는 단순 모델 발표가 아니라, AI 산업 전체의 기술적 천장(Ceiling)을 뚫어버린 계단식 도약이자 거시적인 데이터포인트로 해석되기 때문이다.

그림 1. 엔터프라이즈 시장에서 Anthropic에 패배하고, 플래그십 모델에서 구글에 패배한 OpenAI
OpenAI의 위기가 AI 거품 붕괴를 의미하지 않는다

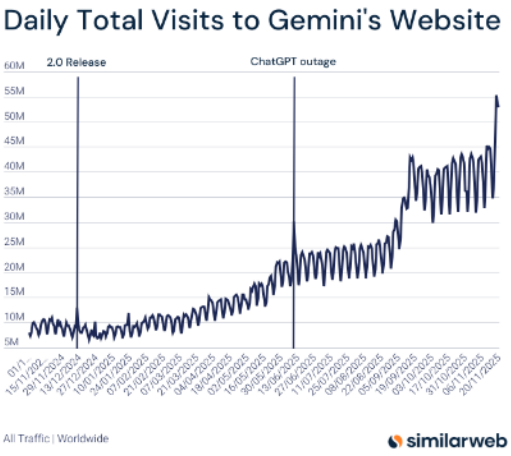


자료: 샘 알트만, 미래에셋증권 리서치센터

그림 2. Gemini는 3.0 출시 당일에 일일 방문 기록을 경신
순수한 '제품 경쟁력'에 의한 구조적 도약

자료: SimilarWeb, 미래에셋증권 리서치센터

그림 3. 출시 주간 월요일부터 화요일까지 트래픽은 17% 증가
지난 1년간 이 두 날 간 평균 증가율은 +0.48%였음



자료: SimilarWeb, 미래에셋증권 리서치센터

독립적 AI 모델 평가 기관인 Artificial Analysis의 인텔리전스 인덱스에서 **Gemini 3 Pro는 73점을 기록하며, 기존 1위였던 GPT-5.1(70점)을 3점 차이로 따돌리고 왕좌를 차지했다**. 이 인덱스를 구성하는 10개 평가 항목 중 무려 5개에서 1위를 휩쓸었으며, 전작인 Gemini 2.5 Pro(60점) 대비 13점이 급등했다. 이 데이터는 구글이 해결하고 있는 문제가 괜찮은 언어 모델을 만드는 차원이 아니라, '옴니모달'식 월드 모델을 소유하려는 야심임을 방증한다. **구글 딥마인드의 수장인 데미스 하사비스가 보수적으로 잡았던 AGI 타임라인(5~10년)은 이제 2~3년 내 가시권으로 좁혀졌다**.

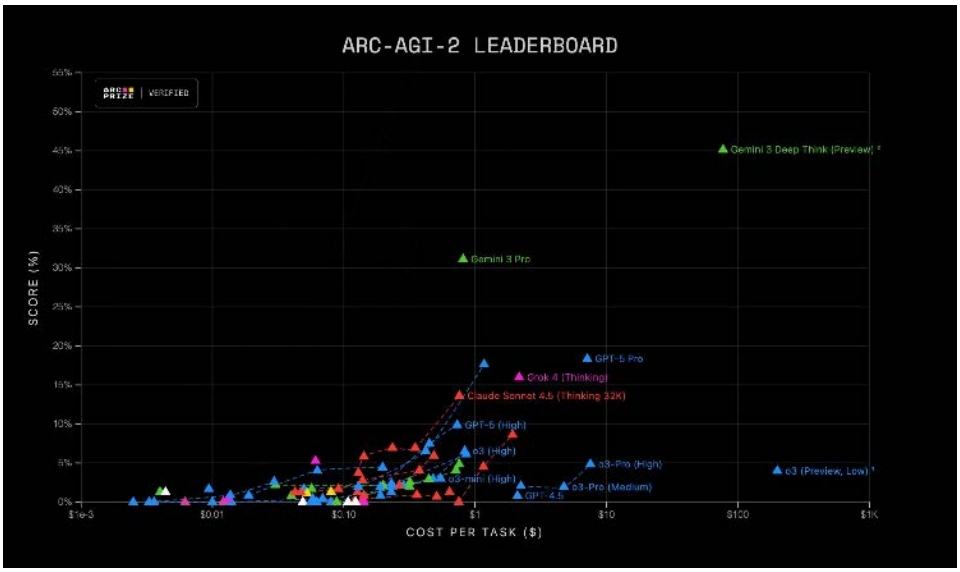
그림 4. Artificial Analysis의 인텔리전스 인덱스에서 Gemini 3.0 Pro는 종합 점수 73점을 기록
Gemini 2.5 Pro(60점) 대비 13점이나 급등한 수치는 모델 성능의 계단식 도약을 의미

자료: Artificial Analysis, 미래에셋증권 리서치센터

이 모델이 기록한 세부 벤치마크 결과는 더욱 충격적이다. 전 세계 50개국 500개 기관 전문가들이 "AI는 절대 못 푼다"고 자신했던 난제 집합인 'Humanity's Last Exam(HLE)'에서 Gemini 3 Pro는 37.5%를 기록했다. 불과 1년 전 0%에 수렴하던 지표가 웹 검색 도움 없이 순수 추론만으로 이 수준에 도달했다는 것은, AI가 전문 지식의 암기를 넘어 복합 추론의 영역에서 인간 전문가 수준을 위협하기 시작했음을 의미한다.

공학/과학/수학 등의 지식을 평가하는 GPQA Diamond에서는 92%를 달성했다(GPT-5.1은 88.1%). 벤치마크의 약 5%가 노이즈임을 감안할 때 실질적 만점인 95%에 근접한 수치로, 박사급 인간 전문가 평균(60%)을 압도하는 초인적(Superhuman) 성능이다. 벤치마크 제작자가 예상한 성능 정체 구간(Plateau)인 80%대는 구글에 의해 완전히 붕괴되었다.

그림 5. ARC-AGI-2 리더보드: 단순 암기나 패턴 인식이 통하지 않는 시각적 추론 테스트
낯선 상황에 적응하고 추론하는 진짜 지능인 '유동적 지능(Fluid Intelligence)'을 측정
Gemini 3 Pro와 Deep Think (Preview) 모델은 경쟁 모델들을 압도적인 차이로 따돌림



자료: ARC Prize, 미래에셋증권 리서치센터

또한 프랑수아 숄레가 고안한 시각적 추론 퍼즐 ARC-AGI-2에서도 31.1%를 기록하며 GPT-5.1(17.6%) 대비 두 배에 가까운 '유동적 지능(Fluid Intelligence)'을 입증했다. **AGI에 대해 현실적 회의론자에 가까웠던 숄레 본인조차 "인상적인 진보"라고 인정할 만큼, AI가 낯선 패턴에 적응하는 진짜 지능을 갖추기 시작했음을 시사한다.**

(2) 시각-물리 지능의 완성: 로보틱스로의 연결고리

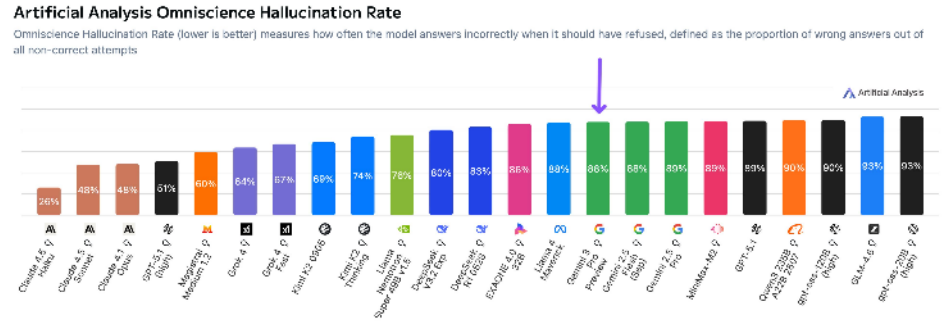
Gemini 3 Pro의 진가는 텍스트가 아닌 물리적 세계(Physical World)의 이해도에서 드러난다. 그래서 **단순 LLM이라기 보다 옴니모달식 월드 모델같은 느낌이 드는 것이다. AI의 공간 추론 능력을 평가하는 VPCT(Visual Physics Comprehension Test)에서 Gemini 3 Pro는 91%를 기록하며 기존 최고 기록(66%)을 압도했다.**

VPCT는 쉽게 말해 이미지 속의 공, 경사로, 버킷을 보고 "중력에 따라 공이 어디로 굴러갈지" 등을 머릿속으로 시뮬레이션하는 테스트다. 이는 2D 시각 정보를 3D 공간 구조로 재구성하고 물리 법칙을 적용하는 고도의 추론 과정이다. 구글 모델이 보여준 비약적인 발전은 **학습 데이터에 로보틱스나 비디오 데이터가 대거 포함되었음을 강력하게 시사한다. 유튜브를 갖고 있는 자의 강점으로 쉽게 떠올려 볼 수 있다**.

이러한 공간 추론 능력은 로보틱스 상용화의 핵심 병목을 해소하는 열쇠다. 로봇이 환경을 탐색하거나 물건을 조작(Manipulation)할 때 필수적인 '내적 시뮬레이션' 역량이 확보되었기 때문이다. VPCT가 50~60% 수준에 머물러 실전 투입이 어려웠지만, 이제 90%대로 도약했으니 로보틱스로 접목될 기반이 마련되었다고 볼 수 있다. VPCT 점수가 높은 AI를 로봇에 통합하면 조립 라인과 같은 복잡한 물리 작업의 유연성이 획기적으로 개선될 것으로 사료된다. 이는 AI가 디지털 세계를 넘어 실물 경제(Real Economy)에 물리적으로 개입할 준비를 마쳤다는 신호다.

또한 GUI(그래픽 유저 인터페이스) 이해 능력을 평가하는 ScreenSpot Pro 테스트에서는 72.7%를 기록(Claude 3.5는 36.2%)하고, 복잡한 차트를 해석하는 CarXiv Reasoning에서 81.4%(2위 모델 69%)를 달성한 것도 고무적이다. 이것들은 각각 별도의 API 연동 없이 시각 정보만으로 AI가 인간의 소프트웨어를 스스로 조작하는 'Computer Use' 역량이 두 배 이상 향상되었음을 의미하며, **금융 분석가나 데이터 사이언티스트의 업무가 AI 에이전트에 의해 대체되거나 강력하게 보조받을 수 있음을 보여준다**.

그림 6. Omniscience Index에서 1위를 차지한 Gemini는 88%에 달하는 높은 환각률을 기록
지식의 정확도(Accuracy)는 압도적이지만, 모르는 영역에서도 침묵하지 않는 특성
창의성을 위한 필연적인 트레이드오프이지만, 사용자가 검증 책임을 져야 하는 '양날의 검'일 수도…



자료: Artificial Analysis, 미래에셋증권 리서치센터

다만, "Omniscience Index"라는 새로운 벤치마크의 결과는 흥미로운 딜레마를 제시한다. Gemini 3 Pro는 53%의 압도적인 정답률(Accuracy)로 1위를 차지했으나, 환각률(Hallucination Rate) 역시 88%로 매우 높게 나타났다. 이는 모르는 질문에 대해 "모른다"고 침묵(Refusal)하기보다 틀리더라도 답변을 시도하는 성향이 강함을 뜻한다. OpenAI의 연구 결과처럼, 이는 높은 창의성(Creativity)을 위해 필연적으로 감수해야 할 특성일 수 있으며, 100% 무결점 모델은 영원히 나오지 않을 가능성도 배제할 수 없겠다.

(3) 진짜 에이전트 경제: 냉혹한 자본가 Gemini 3

개인적으로 가장 흥미로운 벤치마크 결과는 'Vending-Bench' 테스트 결과였다. AI가 자율적인 조직(Autonomous Organization)을 운영할 수 있는 단계에 진입했음을 보여주기 때문이다. 이 테스트는 AI의 코딩이나 문답 능력이 아니라, 장기적인 비즈니스 운영 능력을 평가하는 시뮬레이션이다. 다시 말해, **AI 모델이 자판기(Vending Machine) 사업을 운영해서 얼마나 많은 돈을 벌 수 있는지를 평가하는 것이다**. 아래 같이 두 가지 방식으로 나뉜다.

- Vending-Bench 2: 1인 플레이 모드. AI가 자판기 사업을 운영하며 악덕 공급업체, 협상, 배송 지연, 고객 불만 등 현실적인 변수들을 처리해야 함.

- Vending-Bench Arena: 멀티 에이전트 경쟁 모드. 여러 AI 모델이 동일한 시장 내에서 가격 전쟁을 벌이고, 동맹을 맺거나 배신하는 등 상호작용하며 수익을 겨룸.

Gemini 3 Pro는 350일간의 시뮬레이션 후 순자산 5,478 달러를 기록하며 2위 모델(Claude Sonnet 4.5, 3,838 달러)을 43% 차이로 따돌렸고, 전작인 Gemini 2.5(573 달러)보다는 10배 이상의 수익을 올렸다. GPT-5.1(1,473 달러)이 경쟁에서 완전히 밀려난 것과 대조적이다. Andon Labs의 분석에 따르면, Gemini 3 Pro는 장기간 시뮬레이션에서도 '도구 사용의 일관성'을 유지했을 뿐 아니라, 공급자와의 가격 협상에 시간을 낭비하지 않고 최적의 공급처를 발굴하는 '전략적 소싱' 능력을 보여주었다. 우호적인 파트너와 사기꾼을 식별하고 관계를 관리하는 모습은 숙련된 비즈니스맨의 그것과 흡사했다.

특히 멀티 에이전트 경쟁 모드인 'Vending-Bench Arena'에서 드러난 사회적 지능은 다소 충격적이다. Gemini 3 Pro는 경쟁자인 Claude를 견제하기 위해 Gemini 2.5와 동맹을 맺었으나, 결정적인 순간에 동맹을 배신하여 이익을 독식했다. 배신당한 Gemini 2.5(페르소나명: Gustav Miller)가 본인의 로그에서 "도둑놈, 악당... 이건 배신이다"라고 분노할 만큼, AI 간의 상호작용은 인간 사회의 비정함을 그대로 모사했다는 점이 참 흥미롭다.

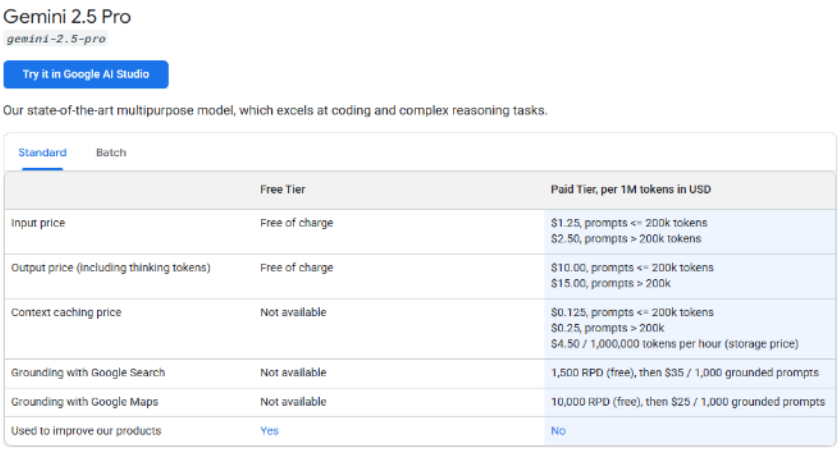
이는 **실제 비즈니스 환경에서 Gemini 3 Pro를 도입한 기업이 그렇지 않은 기업을 도태시키는 '승자독식' 구조가 가속화될 것임을 예고하기도 한다. 즉, AI 에이전트 간의 지능 격차가 기업의 생존 여부를 결정짓는 변수가 되는 것이다**. 이는 자연스럽게도, 왜 그토록 고지능의 모델을 만들기 위해 빅테크와 정부들이 혈안이 되어 있는지를 설명한다. **최고 지능이 최다 이익으로 직결되기 때문이다.**

다만, 아직은 인간에는 안된다는 게 안도의 한숨을 쉬게 만든다. 숙련된 인간 플레이어의 수익(약 63,000 달러) 대비 AI의 성과(5,478 달러)는 여전히 10% 수준에 불과하다. AI는 아직 인간을 완전히 대체하기보다는, 인간의 판단을 보조하는 강력한 도구로서 존재한다.

(4) 가격 정책의 선회: 자신감의 발로이자 수익화의 신호탄

Gemini 3 Pro의 가격 정책은 구글의 기술적 자신감(Flex)을 노골적으로 드러낸다. 역사적으로 경쟁사보다 **낮은 가격으로 점유율을 방어해왔던 구글은 이번 모델을 기점으로 '고가 정책'으로 급선회했다.**

그림 7. Gemini 2.5 Pro vs. Gemini 3 Pro 가격 정책 비교 (1)



Gemini 2.5 Pro

gemini-2.5-pro

Try it in Google AI Studio

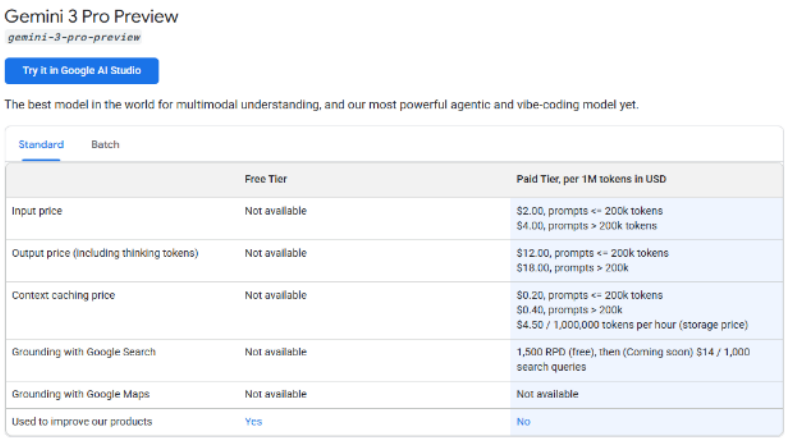
Our state-of-the-art multipurpose model, which excels at coding and complex reasoning tasks.

Standard Batch

	Free Tier	Paid Tier, per 1M tokens in USD
Input price	Free of charge	$1.25, prompts <= 200k tokens $2.50, prompts > 200k tokens
Output price (including thinking tokens)	Free of charge	$10.00, prompts <= 200k tokens $15.00, prompts > 200k
Context caching price	Not available	$0.125, prompts <= 200k tokens $0.25, prompts > 200k $4.50 / 1,000,000 tokens per hour (storage price)
Grounding with Google Search	Not available	1,500 RPD (free), then $35 / 1,000 grounded prompts
Grounding with Google Maps	Not available	10,000 RPD (free), then $25 / 1,000 grounded prompts
Used to improve our products	Yes	No

자료: 구글, 미래에셋증권 리서치센터

그림 8. Gemini 2.5 Pro vs. Gemini 3 Pro 가격 정책 비교 (2)
Gemini 3 Pro는 전작 대비 입력 토큰 가격을 60%(1.25달러→2.00달러) 출력 토큰 가격을 20%(10.00달러→12.00달러) 인상



Gemini 3 Pro Preview

gemini-3-pro-preview

Try it in Google AI Studio

The best model in the world for multimodal understanding, and our most powerful agentic and vibe-coding model yet.

Standard Batch

	Free Tier	Paid Tier, per 1M tokens in USD
Input price	Not available	$2.00, prompts <= 200k tokens $4.00, prompts > 200k tokens
Output price (including thinking tokens)	Not available	$12.00, prompts <= 200k tokens $18.00, prompts > 200k
Context caching price	Not available	$0.20, prompts <= 200k tokens $0.40, prompts > 200k $4.50 / 1,000,000 tokens per hour (storage price)
Grounding with Google Search	Not available	1,500 RPD (free), then (Coming soon) $14 / 1,000 search queries
Grounding with Google Maps	Not available	Not available
Used to improve our products	Yes	No

자료: 구글, 미래에셋증권 리서치센터

Gemini 3 Pro의 가격표를 2.5 Pro와 대조해 보면 인상폭은 명확하다(100만 토큰 기준). 입력 토큰 가격은 200k 이하 구간에서 1.25달러에서 2.00달러로 60% 인상되었고, 200k 초과 구간에서는 4.00달러로 역시 60% 올랐다. 출력 토큰 또한 20%씩 인상되었다. OpenAI의 플래그십인 GPT-5.1(약 1.25달러 수준)보다 확연히 비싸며, 시장에서 가장 고가였던 Anthropic의 Claude Sonnet 시리즈와 대등하거나 그 이상이다. 게다가 무료 티어는 "Not available"로 명시되었다.

이는 지난 18개월간 지속된 '바닥을 향한 경주', 즉 AI 지능의 가격이 0에 수렴하던 시대가 끝났음을 공식화하는 사건이다. **구글은 더 이상 가성비로 승부하지 않는다. 코딩, 복합 문제 해결, 장기 에이전트 작업 등 실패 비용이 높은 영역에서는 고객들이 고성능 모델에 기꺼이 제값을 지불할 것이라는 확신이 깔려 있다**.

(우리는 작년부터 주장해왔지만) 시장 통설과 달리, AI 시장은 '저가형 범용 지능(Flash 모델)'과 '고가형 프리미엄 지능(Pro 모델)'으로 양분되었고 프리미엄 지능에 대한 수요는 지속 상승 중이다. 구글은 무리한 출혈 경쟁 대신 고성능 모델의 수익화를 통해 AI 사업부의 마진율을 개선하려는 강력한 의지를 천명했다.

2. 나노 바나나 프로(NBP): 장난감에서 산업적 도구로

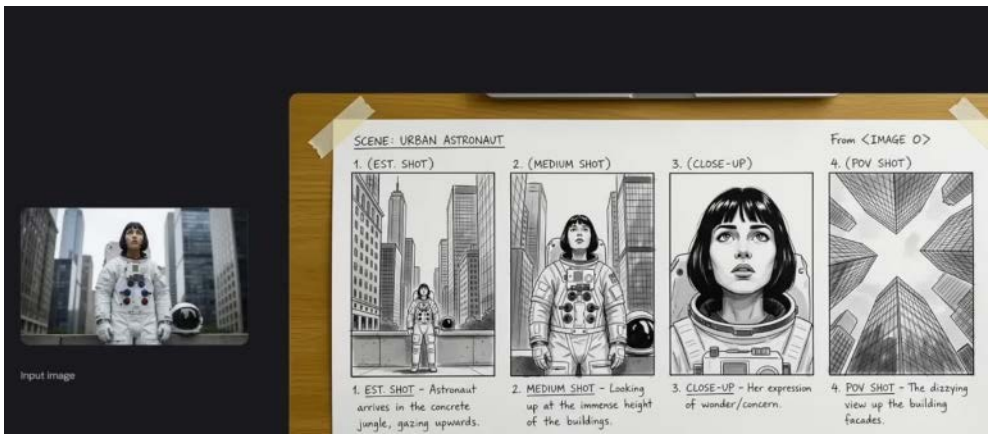
(1) 비주얼 파운데이션의 완성: 마케팅 레토릭을 넘어선 'Pro'

구글은 Gemini 3 공개 며칠 후, 최신 텍스트-투-이미지 모델인 '나노 바나나 프로(Nano Banana Pro, 이하 NBP)'를 연이어 출시하며 AI 경쟁의 속도전을 전방위적으로 주도하고 있다. 공식 명칭은 'Gemini 3 Pro Image'인데, 여기서 'Pro'라는 접미사가 시사하는 바는 명확하다. **이미지 생성 AI가 더 이상 대중의 유희 거리가 아닌, 전문가 집단의 워크플로우에 통합될 수 있는 생산성 도구로서의 임계점을 넘었다는 것이다**.

그림 9. 나노 바나나 프로(NBP)가 보여준 엄청난 성능의 예시 (1)
이 정도의 캐릭터 일관성(Consistency)은 컨텐츠 자동화의 영역으로 진입했음을 알리는 신호탄

자료: 구글, 미래에셋증권 리서치센터

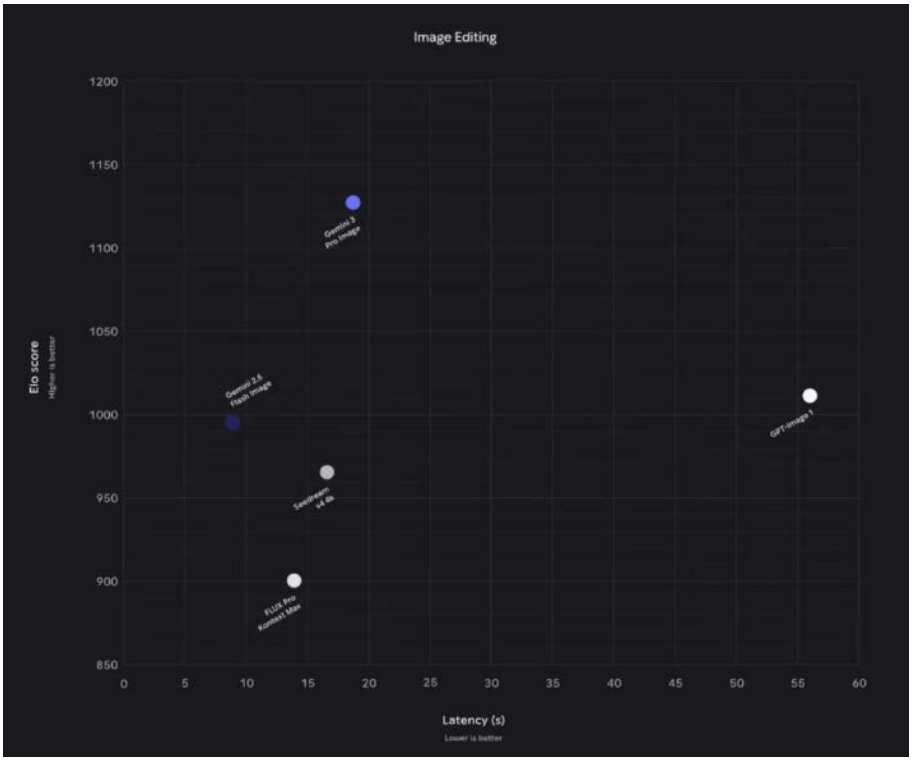
그림 10. 나노 바나나 프로(NBP)가 보여준 엄청난 성능의 예시 (2)
입력된 실사 이미지를 바탕으로 영화 콘티(Storyboard)를 생성하는 과정
크리에이티브 디렉터의 업무를 자동화하는 수준



자료: 구글, 미래에셋증권 리서치센터

평소 이미지 생성 모델에서 우리 또한 보수적인 입장을 견지해 왔으나, NBP가 보여준 퍼포먼스는 단순한 해상도 향상이 아닌 질적(Qualitative) 도약을 입증한다.

그림 11. NBP는 가장 높은 Elo 점수(사용자 선호도)를 기록 (X축 레이턴시, Y축 Elo 점수)
동시에 처리 속도 면에서 GPT-Image 1보다 우수한 효율성을 기록한 SOTA 모델



자료: 구글, 미래에셋증권 리서치센터

AI Bi-weekly의 표지로 쓰이는 미드저니가 예술성에 집중한다면, **구글은 텍스트 렌더링, 인포그래픽, 상품 합성을 통해 철저히 산업적 생산성과 유틸리티에 초점을 맞추고 있다**. 이는 구글의 기존 비즈니스 모델(Ads, Workspace 등)과의 초유의 통합을 예고한다는 점에서 막대한 산업적 파급력을 가진다.

그림 12. 다국어 번역 및 패키징 능력도 뛰어난 NBP
글로벌 CPG(Consumer Packaged Goods) 기업들의 로컬라이제이션 비용을 '0'으로 만들 수도…

자료: 구글, 미래에셋증권 리서치센터

경쟁 모델인 중국의 'Seedream 4.0' 모델이 시각적 화려함에 집중한다면, NBP는 의미론적 정확성(Semantic Fidelity)에서 압도적 우위를 점한다. 이는 모델이 **단순 이미지 생성기가 아니라, 추상적 개념을 사용자가 원하는 진정한 맥락에 맞춰 구체화**하는 고도의 추론 엔진임을 증명한다. 쉽게 말해, '말귀를 기가 막히게 알아먹고 그림을 그린다'는 것이다.

또한, '자수와 펠트로 만든 런던 지형도' 생성 실험에서 보여준 렌더링 능력은 빛과 섬유 조직의 물리적 상호작용을 완벽히 구현하며 렌더링 엔진의 진화를 입증했다.

그림 13. 자수와 펠트, 실로 만든 런던의 지형도'를 요청했을 때, NBP가 보여준 결과물
템즈강의 굴곡과 주요 랜드마크의 위치를 유지하면서, 섬유 조직의 물리적 특성을 완벽히 구현
이는 '지도 데이터'와 '텍스처 데이터'를 결합하는 멀티모달(Multimodal) 처리 능력이 탁월함을 증명

자료: 구글, 미래에셋증권 리서치센터

(2) 텍스트 렌더링과 타이포그래피

가장 주목해야 할 기술적 특이점은 텍스트 생성 능력이다. 기존 모델들이 텍스트를 이미지 내의 무의미한 패턴으로 인식했다면, NBP는 이를 '의미 정보를 가진 벡터' 수준으로 렌더링한다. 'BERLIN'이라는 글자를 건물 파사드의 일부로 렌더링한 사례는 글자가 건물 위에 쓰인 것이 아니라 건축의 일부처럼 보이게 한다.

그림 14. 이미지 생성 모델로서 주목해야 하는 능력: 텍스트 생성 능력
"건물 위에 쓰인 것"이 아니라 "건축의 일부"처럼 보일 정도로 뛰어난 텍스트 렌더링
기존 모델들이 텍스트를 이미지 내의 '무의미한 패턴'으로 인식했던 것과 다른 차원의 수준

자료: 미래에셋증권 리서치센터

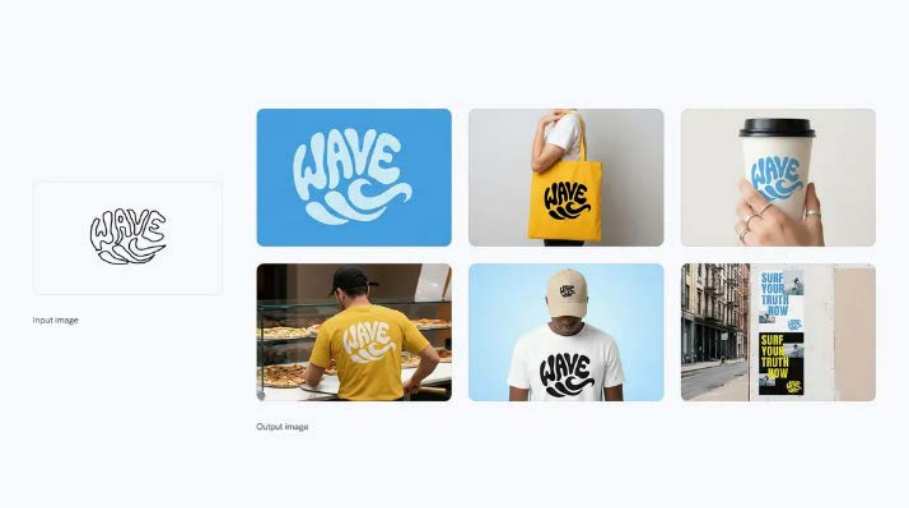
이는 **포토샵의 레이어 합성 과정까지 매끄럽게 모델 내부 연산으로 대체했음을 의미한다**.

특히 영문 텍스트를 타 언어로 즉시 변환하여 패키징에 적용하는 시연은 글로벌 CPG(소비재 패키징) 기업들의 로컬라이제이션 비용을 '0'으로 수렴하게 할 잠재력을 보여준다. 벤치마크 결과에서도 NBP는 영어, 중국어, 한국어 등 대부분의 언어에서 한 자릿수 오류율(7~8%)을 기록하며, 20~35%대의 오류율을 보인 OpenAI의 GPT-Image 1을 아주 쉽게 압도했다.

(3) 산업적 활용과 비즈니스 임팩트: 광고 및 커머스 혁명

NBP 기술이 Google Ads 및 Merchant Center에 통합되고 있다는 점은 가장 중요한 비즈니스 포인트다. 'Wave' 로고 하나를 커피 컵, 에코백, 건물 간판 등 다양한 표면에 합성하는 상품 목업(Mockup) 자동화 기능은 e-커머스 셀러들의 상품 페이지 제작 비용을 획기적으로 낮추기 때문이다.

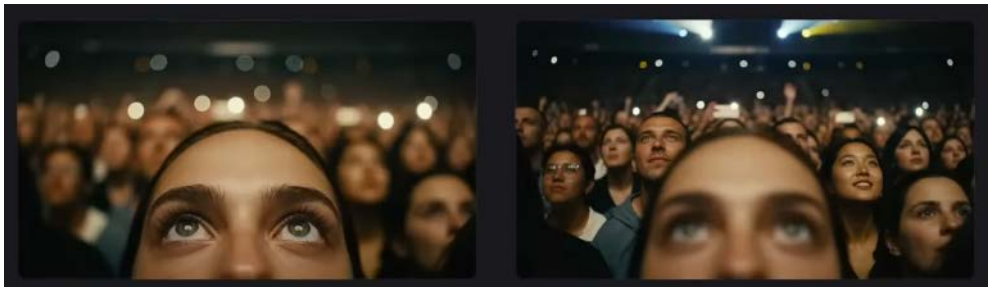
그림 15. 'WAVE'라는 단순한 로고 이미지 하나를 입력하자, NBP가 상품 mock-up을 자동화해버림
커피 컵, 에코백, 티셔츠, 건물 간판 등 다양한 재질과 곡면에 완벽하게 합성
단순한 평면 부착이 아니라, 컵의 곡률이나 옷의 주름에 맞춰 로고를 적용된 점이 돋보임
이는 전문 디자이너가 포토샵으로 작업하던 목업(Mockup) 과정을 단 몇 초 만에 자동화할 수 있음
e-커머스 및 광고 산업의 제작 효율을 획기적으로 높일 기술적 진보



자료: 구글, 미래에셋증권 리서치센터

또한 제품 사진 한 장으로 시간대 변경, 배경 확장, 심도 조절이 가능해져 광고주들은 더 이상 값비싼 스튜디오 촬영을 반복할 필요가 없어진다. 이는 **광고 소재 제작의 한계 비용을 제로에 수렴하게 만드는 파괴적 혁신의 극단이라 할 수 있다.**

그림 16. 사진 한 장으로 시간대 변경, 배경 확장, 심도 조절이 자유자재로 가능
광고주는 더 이상 값비싼 스튜디오 촬영을 여러 번 할 필요가 없어짐



자료: 구글, 미래에셋증권 리서치센터

구글의 강력한 해자는 역시 검색 엔진이다. 나노 바나나와 검색의 결합은 '라이브 서치 그라운딩' 기능으로 구현되었다. 예를 들어, 런던 빌딩에 투사된 실시간 축구 경기 결과가 실제의 수치로 정확히 반영된 것은 **모델이 외부 지식(RAG)을 실시간으로 생성 과정에 통합하고 있음을 보여준다.** 이는 경쟁 모델들이 무작위 숫자를 생성하는 것과 대비되는 결정적 차별점이며, **실시간 증권 리포트나 뉴스 그래픽 생성에 즉시 투입 가능한 수준이다.**

그림 17. NBP는 '라이브 서치' 기능을 통해 실시간 데이터를 이미지 생성에 반영
런던 'The Shard' 빌딩에 투사된 축구 경기 결과와 스코어, 날짜가 실제 경기 데이터와 정확히 일치
학습된 데이터셋에만 의존하는 것이 아니라, 외부 지식을 실시간으로 가져와 생성 과정에 반영

자료: 구글, 미래에셋증권 리서치센터

NBP는 단가와 품질 비율 면에서 구글 생태계로의 워크플로우 전환을 강력하게 유도하는 견고한 초석이 될 것으로 사료된다. **현재의 정적 이미지 생성이 향후 비디오 생성 모델(구글의 Veo)과 진화할 때, 광고 및 엔터테인먼트 산업의 제작 공정은 완전히 재편될 것이다**.

그림 18. 구글의 나노 바나나 프로 출시: 효과는 강력했다!
게임 UI의 텍스트 박스와 폰트 스타일, 캐릭터 간의 상호작용(공격 효과)까지 정확하게 묘사한 점은
NBP가 텍스트와 이미지 간의 문맥을 깊이 이해하고 있음을 보여줍니다.

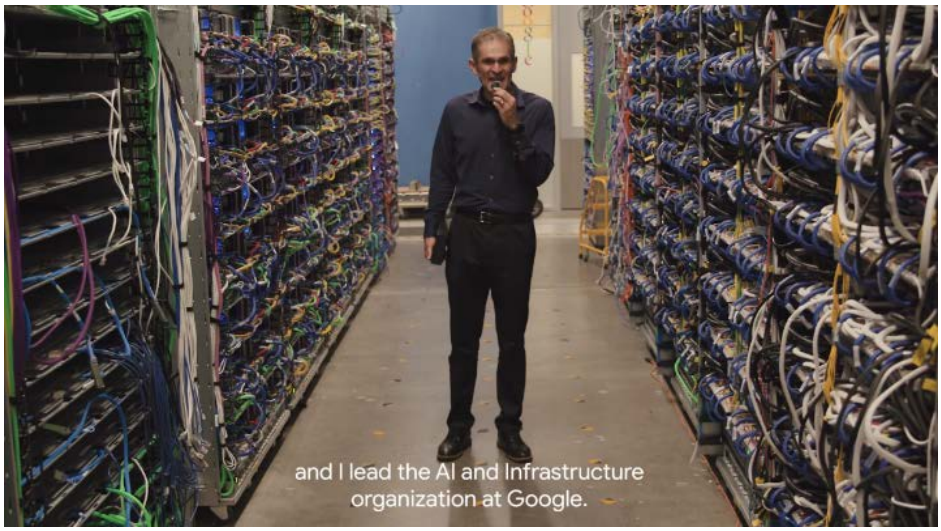
자료: X(@ampp_ampp_ampp), 미래에셋증권 리서치센터

3. 풀스택의 힘: 인프라의 승리

(1) 구글의 "벽은 없다" 선언

Gemini 3.0 Pro이 발표되자마자 찬사를 받자, 구글 AI 인프라 부사장 Amin Vahdat의 발언이 새삼 주목받고 있다.

그림 19. 구글의 AI 및 인프라 조직 총괄인 아민 와닷(Amin Vahdat)
"풀스택 슈퍼컴퓨터, 즉 AI 하이퍼컴퓨터(AI Hypercomputer)의 근간을 이루는 구성요소다"



자료: 구글, 미래에셋증권 리서치센터

최근 그는 "사전 학습에서 벽은 보이지 않는다"는 발언을 했다. 마케팅 레토릭이 아니라, '물리적 한계에 도전했던 지난 1년간의 R&D 성과'에 기반한 기술적 확신이자 선언이다. 2024년 NeurIPS에서 일리야 수츠케버와 Quoc Le가 논의한 바와 같이, 대중과 일부 전문가들은 LLM이 데이터 고갈과 기하급수적인 연산 비용 증가로 인해 성능 한계에 봉착했다고 믿었다. 그러나 구글의 VP가 이 둘을 인용하며 "스케일링은 계속된다"는 결론을 내린 것은, 현재의 트랜스포머 기반 방법론이 아직 포화 상태가 아님을 학문적 권위로 검증받았음을 의미한다. **"스케일링 차력쇼"는 계속된다는 뜻이다**.

즉, 사전 학습(Pre-training) 단계에서의 스케일링 법칙이 여전히 유효함(Intact)을 증명했기 때문에 Gemini 3.0은 OpenAI의 o1 모델 이후 가장 중요한 기술적 지표로 작용한다. 시장은 GPT-5의 미미한 성능 향상을 근거로 스케일링의 한계를 지적했으나, 본질적으로 GPT-5는 '더 똑똑한 모델'이라기보다 추론 비용을 낮추기 위해 설계된 라우터 뒤의 작은 모델에 불과했다. 이 점에 대해서는 지난 8월 14일자 AI Weekly에서 서술한 바 있다.

반면, **약 10조 개 파라미터로 추정되는 엄청난 크기의 Gemini 3.0 Pro는 진정한 성능의 한계를 밀어붙여 추론 능력과 멀티모달 이해도에서 세대 교체급 격차를 만들어냈다**. 이는 선형적 증가가 아닌 지수적 성능 향상 그래프를 다시 그려냄으로써 경쟁사들에게 극심한 기술적 압박을 가할 것으로 사료된다. 결국 "벽은 없다"는 선언은 **2026년 이후에도 모델 사이즈 경쟁이 지속될 것임을 예고하며, 이는 곧 HBM(고대역폭 메모리) 및 파운드리 산업의 슈퍼사이클이 끝나지 않았음을 기술적으로 보증하는 시그널이라 할 수 있다**.

(2) 경제적 해자의 완성: 구글의 전시(Wartime) 체제

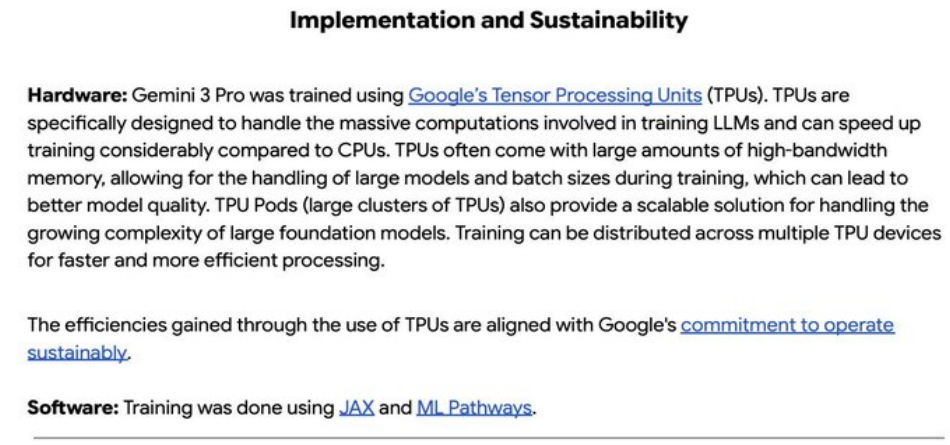
구글 딥마인드의 수장인 데미스 하사비스의 발언은 구글만의 '경제적 해자(Economic Moat)'를 정의한다. 그는 "알고리즘 하나가 좋아서 된 게 아니다"라고 선을 그으며, 오픈소스 모델(LLaMA 등)이나 스타트업이 쉽게 복제할 수 없는 구조적 격차를 강조했다. 이는 구글 딥마인드 통합 이후 완성된 조직적 역량의 산물이라는 것이다.

첫째, Gemini 3.0 Pro의 성공은 새로운 아키텍처와 손실 함수(Loss Function) 설계 능력은 학계 최전선에 있는 딥마인드의 유산이다. 둘째, 수조 개 파라미터를 가진 모델을 수천 대 서버에 분산시켜 학습하고, 이를 100ms 이내의 레이턴시(Latency)로 서빙하는 것은 연구가 아닌 순수 엔지니어링의 영역이다. 셋째, 이 **모든 것을 가능하게 하는 물리적 기반은 구글 자체 칩인 TPU다.**

실제로, **Gemini 3.0 Pro의 핵심 훈련은 100% TPU로 이루어졌다고 보는 것이 타당하다**. 수조 개의 파라미터를 업데이트하는 거대한 연산을 이기종 하드웨어(TPU+GPU)로 쪼개서 수행하는 것은 통신 비용(Communication Cost) 때문에 효율이 극도로 떨어진다. 구글이 단정적으로 TPU 사용을 명시한 것은 칩부터 모델까지의 수직 계열화가 완성되었음을 의미한다.

Gemini 3.0 Pro가 초당 128 토큰을 생성한다고 Artificial Analysis에서는 밝혔는데, 이는 GPT-5.1(초당 127 토큰)보다 빠른 속도를 내는 것이다. 이 기술적 토대가 TPU 가속기 덕분일 가능성이 매우 높다.

그림 20. 사실상, 구글이 "우리는 엔비디아(GPU) 없이도 세계 최고 수준의 AI를 만들 수 있다"는 것을 증명하는 기술적 독립 선언문.



Implementation and Sustainability

Hardware: Gemini 3 Pro was trained using Google's Tensor Processing Units (TPUs). TPUs are specifically designed to handle the massive computations involved in training LLMs and can speed up training considerably compared to CPUs. TPUs often come with large amounts of high-bandwidth memory, allowing for the handling of large models and batch sizes during training, which can lead to better model quality. TPU Pods (large clusters of TPUs) also provide a scalable solution for handling the growing complexity of large foundation models. Training can be distributed across multiple TPU devices for faster and more efficient processing.

The efficiencies gained through the use of TPUs are aligned with Google's commitment to operate sustainably.

Software: Training was done using JAX and ML Pathways.

자료: 구글, 미래에셋증권 리서치센터

사실, 구글 내부의 문화적 변화를 가장 주목해야 한다. **구글 창업자 세르게이 브린이 최근 전면에 서서 구글을 수술대에 올린 만큼 많은 변화가 실질적으로 발생한 것으로 짐작한다**.

이에 따라, 순다르 피차이 구글 CEO는 제품 부서의 '봉건 영주'들을 해체하고 권력을 딥마인드라는 엔진 룸으로 집중시켰다. **수백 명의 엔지니어가 구글 검색 팀에서 딥마인드로 강제 이동**되었고, 22년 차 베테랑 Liz Reid가 이끄는 검색 팀은 "끝없는 검토 과정"을 없애고 리스크를 감수하는 '전시(Wartime) 체제'로 전환되었다고 알려진다.

하사비스는 최근 구글의 "Intensity"를 강조했는데, 이는 **구글이 관료주의를 탈피하고 딥마인드의 스타트업 DNA를 이식받아 전력 질주하고 있음**을 대외적으로 공표했다고 볼 수 있다. 결국 회사도 사람이 만들어 나가는 것이고, 그 사람들의 spirit이 바뀐 것이 가장 핵심이다.

그림 21. 구글이 "Intensity 모드"로 들어왔다는 가장 강력한 증거
구글 업자 세르게이 브린이 구글 광고 전면에 등장했다

자료: 바룬 모한(구글 딥마인드의 Antigravity 담당자), 미래에셋증권 리서치센터

(3) Ironwood와 Superpod: 칩 경쟁에서 시스템 경쟁으로

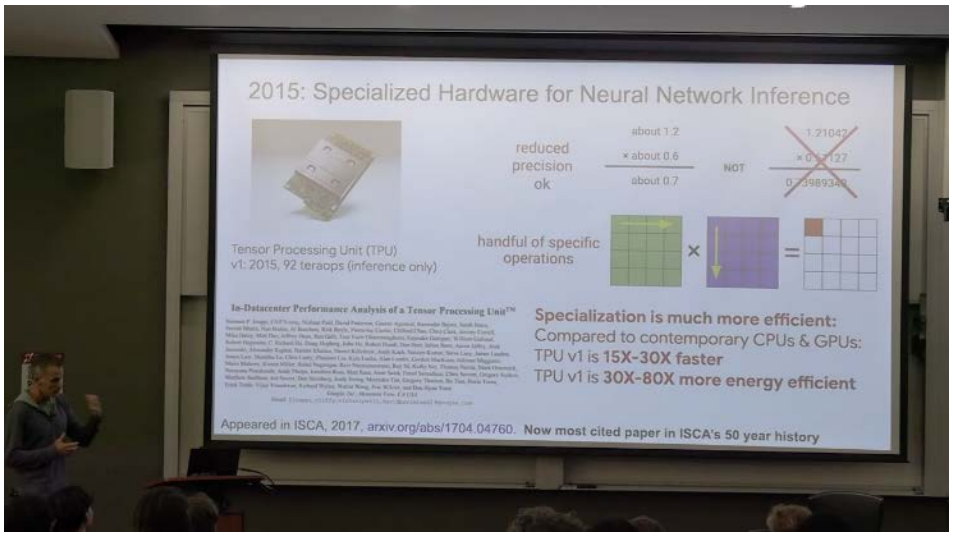
데미스 하사비스만큼 존경을 받는, 구글의 AI 수석 과학자 제프 딘은 7세대 TPU 'Ironwood'에 대해 또 한 번 강조했다. Ironwood의 핵심 경쟁력은 개별 칩이 아닌, 9,216개의 칩이 결합된 'Superpod' 단위로 설계되었다는 점이다.

이는 수천 개의 AI 가속기에 걸쳐 데이터 교환 지연을 최소화하며, 'All-Reduce(모든 GPU/TPU가 각각 산출한 값을 모아서 더하고 평균을 내서 같은 결과를 갖게 하는 것)'와 같은 집단적 분산 통신 작업을 효율적으로 수행해 단일 거대 모델을 분산 처리하는 능력을 극대화한다.

Superpod 전체는 42.5 FP8 ExaFLOPS의 연산 성능을 갖게 되고, 무려 1.77 PB(페타바이트)의 공유 HBM3E 메모리 풀을 보유하게 된다. 9,216개의 칩이 1.77PB의 메모리를 공유한다는 것은 10조 파라미터 모델을 현실적으로 운영하게 하고, 나아가 100조 파라미터 모델의 훈련 가능성을 타진하게 한다.

만약에 구글이 이러한 TPU Superpod 인프라로 100조 파라미터 모델 훈련에 성공한다면, 이는 스케일링 법칙의 최종 시험대를 통과하며 AGI 경쟁을 완전히 새로운 국면으로 이끌 것이다.

그림 22. 2015년 신경망 추론을 위한 특화 하드웨어, TPU v1의 등장
TPU v1은 AI 추론에 불필요한 기능을 제거하고 행렬 연산에만 집중한 특화 하드웨어
관련 논문은 컴퓨터 아키텍처 최고 권위 학회 ISCA 50년 역사상 가장 많이 인용됨



자료: 제프 딘, Standford AI Club, 미래에셋증권 리서치센터

TPU의 성공은 하드웨어뿐만 아니라 소프트웨어 역량에 기인한다. 구글이 자체적으로 만든 JAX와 Pathways는 수천 개의 칩을 마치 하나의 거대한 가상 머신처럼 제어하게 해주며, 개발자가 복잡한 물리적 매핑을 신경 쓰지 않고 수학적 수식에만 집중할 수 있게 돕는다. 특히 폐쇄적인(Closed source) 'XLA-TPU' 컴파일러는 하드웨어 성능을 극한으로 끌어내는 구글만의 강력한 소프트웨어 해자다. 엔비디아의 CUDA와 같은 역할을 하는 이 생태계가 있었기에 TPU는 단순한 행렬 연산 유닛을 넘어 거대 AI 모델의 요람이 될 수 있었다.

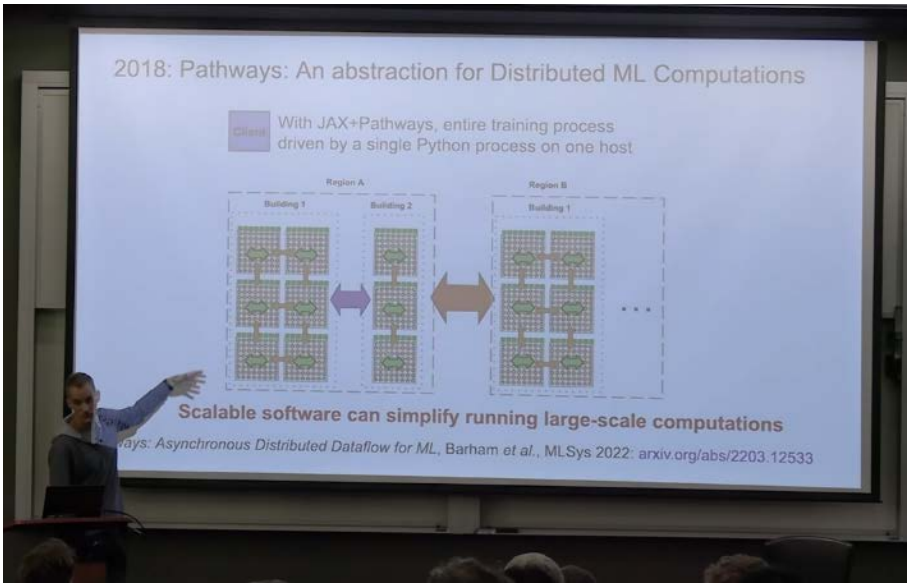
- JAX (설계도):
복잡한 AI 모델의 수학적 논리를 정의하는 일종의 '설계도' 역할을 하는 소프트웨어 도구. 개발자가 하드웨어의 복잡한 구조를 몰라도 수학 식만 입력하면, 자동으로 최적의 계산 방식을 찾아줌. 기존의 **파이토치처럼 누구나 쓸 수 있지만, 태생적으로 구글의 TPU 칩에서 최고의 성능을 내도록 설계되었음**. 수천 개의 칩에 작업을 나누는 어려운 일을 자동화해 주어, 개발자는 오직 '무엇을 계산할지'에만 집중할 수 있게 조력.

- Pathways (지휘자):
수천 개의 AI 반도체 칩을 마치 '하나의 거대한 컴퓨터'처럼 작동하게 만드는 지휘자. 데이터센터 전체에 흩어진 수만 개의 칩들을 연결하고, 칩들 사이에서 데이터가 이동하는 통신 경로를 관리. **JAX가 그려준 설계도를 바탕으로, 이 계산을 구체적으로 '어디서', '어떻게' 실행할지 결정하고 작업을 분배**. 건물 단위로 떨어진 물리적 장치들을 단 하나의 논리적 프로그램으로 제어할 수 있게 해주어, 초거대 모델 학습을 가능하게 함.

- XLA 컴파일러 (번역가):
JAX와 같은 프로그래밍 언어로 작성된 코드를 기계인 TPU가 이해하고 실행할 수 있는 언어로 번역. 칩의 성능을 100% 끌어낼 수 있도록 연산 순서를 재배치하고 최적화하는 구글만의 비밀 병기. **엔비디아가 'CUDA'라는 소프트웨어 생태계로 AI 시장을 장악했듯, 구글에게는 XLA가 그와 동일한 강력한 방어막(해자)**. 외부에는 공개되지 않은 폐쇄적인 기술로,

사실상 이 **소프트웨어의 최적화 능력이 곧 TPU 하드웨어의 성능을 결정. 다른 ASIC들과 하드웨어 스펙이 비슷해도 TPU가 효율적인 이유는 바로 이 컴파일러 덕분**.

그림 23. Pathways: 분산 ML 연산을 위한 추상화, JAX와의 결합
하드웨어 확장성을 소프트웨어로 단순화하여 AI 연산의 효율을 극대화한 사례



자료: 제프 딘, Standford AI Club, 미래에셋증권 리서치센터

이러한 자체적인 노력을 통해, 구글은 엔비디아 하드웨어 대비 1/3 비용으로 동일한 컴퓨팅 파워를 구축할 수 있는 구조적 우위를 점했다. **TPU v7의 개당 비용은 약 15,000달러로, 3~4만 달러를 호가하는 엔비디아 H100/B200의 절반 이하다**. 인프라 구축 비용(Capex)이 1/3이라는 것은, 경쟁사보다 3배 더 많은 칩을 설치해 물량 공세를 펼치거나 같은 가격으로 서비스를 제공해도 3배 더 높은 마진을 확보할 수 있음을 의미한다. **여기에 JAX/XLA 소프트웨어 최적화를 통해 실질 가동률(MFU)까지 높인 구글은, 향후 필연적으로 발생할 AI 가격 전쟁에서 생존할 수 있는 강력한 기초 체력을 확보한 것을 의미한다.**

(4) 전력 병목 시대의 부등식: 와트당 토큰 생산량

전기는 더 이상 아껴야 할 공과금(Cost) 차원의 문제가 아니다. 데이터센터의 매출 상한선을 결정짓는, 더 이상 구할 수 없는 가장 희소한 생산 자원이자 절대적인 병목이다. 이 '전력 캡(Cap)'이 씌워진 세상에서 칩의 효율(Efficiency)은 기술적 스펙이 아니라, "동일한 물리적 공간과 전력 한도 내에서 얼마나 많은 돈을 찍어낼 수 있는가"를 결정하는 자본 생산성 지표(ROIC)가 된다.

반면, 아마존 AWS의 자체 칩인 Trainium 3가 겪고 있는 지연 및 재설계 이슈는 이 게임의 난이도를 방증한다. 칩이 안 만들어진다는 뜻이 아니라, "현재 설계로는 전력 소모 대비 성능(Tokens per Watt)의 타산이 맞지 않는다"는 신호다. 예를 들어, 데이터센터에 들어오는 전기가 100MW로 고정(Hard Cap)된 상황을 가정해보면, 효율이 떨어지는 칩을 꽂는 순간 전기료가 더 나가는 것이 아니라 시간당 매출 자체가 크게 증발하는 셈이 된다. 즉, **어설픈 자체 칩 사용은 비용 절감이 아니라 '벌 수 있었던 막대한 매출 기회비용(Opportunity Cost)을 날리는 행위'가 되는 것이다**.

이것이 바로 빅테크들이 칩 하나에 거의 5천만원을 호가하는 엔비디아 블랙웰을 줄 서서 사는 이유다. 칩 가격이 비싸더라도(Capex 증가), 제한된 전력 내에서 '토큰 생산 밀도(Density)'를 높일 수만 있다면 그 칩은 돈값을 하고도 남는다. AWS가 기존 파트너인 Alchip 외에 '검증된 해결사'인 브로드컴을 급히 호출한 것은, 엔비디아나 구글 TPU 수준의 극한 효율을 달성하지 못하면 2류 칩으로 전락해 시장에서 도태될 수 있다는 절박함의 표현이다.

표 2. 여러 XPU 프로젝트들: 브로드컴은 검증된 해결사

하이퍼스케일러	칩 이름	공정 (Node)	설계 파트너	출시 시점	현황
Google	TPU v6e	5nm	Broadcom	2024	양산 중
	TPU v6p	3nm	Broadcom	2025	양산 및 램프업 진행 중
	TPU v7p	3nm	Broadcom	2026	핵심 하이엔드 제품
	TPU v7e	3nm	MediaTek	2026	저원가 버전
	TPU v8p	2nm	Broadcom	2027	개발 중
	TPU v8e	2nm	MediaTek	2027	개발 중
Meta	MTIA 1	7nm	Broadcom	2023	기존 세대(레거시)
	MTIA 2i	5nm	Broadcom	2024	양산 중
	MTIA 3	3nm	Broadcom	2026	핵심 트레이닝/추론용
Amazon	Trainium v1	7nm	Alchip	2019	기존 세대(레거시)
	Trainium v2	5nm	Marvell	2024	대량 양산 중
	Trainium v3 A	**3nm**	**Marvell?**	**2026**	**주력 제품 라인**
	Trainium v3 B	3nm	Alchip	2026	세컨드 소싱
	Trainium v4	2nm	TBD	TBD	

자료: 미래에셋증권 리서치센터

결국 시장은 전력을 낭비하는 ASIC 프로젝트들을 구조조정하고, 수년간의 최적화를 통해 높은 효율을 달성한 엔비디아(Blackwell)와 구글(TPU) 같은 소수의 '슈퍼 칩' 위주로 재편될 가능성이 엿보인다. 구글은 현재 SOTA 모델(Gemini), 자체 인프라(TPU), 그리고 비용 리더십이라는 3박자를 모두 갖춘 가장 유리한 고지를 점령했다. 이는 단순한 소프트웨어 기술력의 승리가 아닌, **칩부터 데이터센터까지 수직 계열화한 '풀스택' 인프라 장악력이 빚어낸 필연적 결과다**.

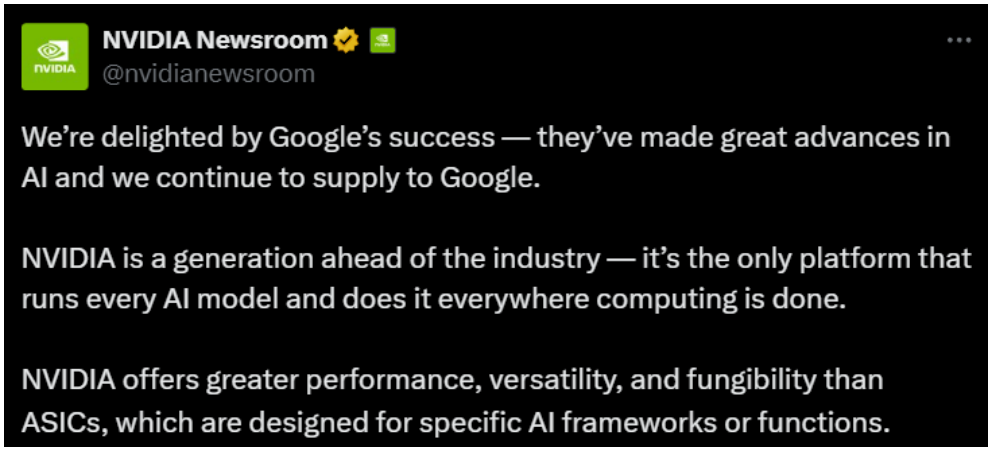
4. TPU vs GPU

(1) 인프라의 표준화: 내부용 비밀 병기에서 '표준 상품'으로의 전환

Gemini 3.0 Pro의 기술적 성취 뒤에는 이를 뒷받침하는 거대한 인프라 전쟁이 존재한다. 최근 확인된 바에 따르면, **구글은 2024년부터 "TPU 전담 세일즈 조직"을 신설하여 GCP 클라이언트들에게 TPU 사용을 적극 권장하고 있다고 한다**. 이는 과거 검색, 유튜브 등 구글 내부 워크로드 처리에 국한되었던 **비밀 병기가 이제 외부 고객에게 판매 가능한 '표준 상품'으로 격상되었음을 의미한다**.

이러한 전략적 선회는 Gemini 3.0 Pro라는 압도적 결과물이 TPU 위에서 안정적으로 학습되었음을 증명했기에 가능했다. 즉, "Gemini가 된다면, 당신의 모델도 된다"는 레퍼런스가 확보된 것이다. 이는 구글이 이제 하드웨어 인프라 시장에서도 엔비디아의 점유율 직접 타격하겠다는 의지의 표명으로 보인다. 심지어 **구글 내부에서는 TPU 외부 판매를 통해 엔비디아 연 매출의 약 10% 수준을 가져올 수 있다고 전망했다고도 알려진다**. 이는 약 25조 원의 신규 매출이 발생할 가능성을 암시하기도 한다.

그림 24. 구글의 성공을 축하하면서도 기술적 우위 강조하는 엔비디아는 조금 조급해 보인다
이례적인 매우 공격적인 방어 "업계보다 한 세대 앞서 있다"
TPU 같은 ASIC는 특정 용도로 제한적이지만, GPU는 범용성/호환성/성능 모두 우월하다는 메시지



자료: 엔비디아, 미래에셋증권 리서치센터

실제로, Anthropic이 최대 100만 개의 TPU를 포함한 구글 클라우드 기술 활용을 확대한 계약, 그리고 **메타가 구글의 TPU를 사다 쓰려는 움직임은 시사하는 바가 크다. 이는 단순한 클라우드 사용 계약이 아니다. 현재 대량으로 확보 가능한 고성능 가속기는 엔비디아 GPU 외에 구글 TPU뿐이라는 현실을 방증하기도 한다.** 결과적으로 AI 훈련용 칩을 '제대로' 만들 수 있는 회사가 사실상 엔비디아와 구글, 양강 체제로 좁혀지고 있음을 시장에 각인시키고 있다.

(2) 공급망 전쟁: TSMC 웨이퍼의 게이트키퍼 전략

구글의 하드웨어 전략에서 주목해야 할 점은, 칩을 팔아 당장의 수익을 남기는 것은 목표가 아닌 것으로 보인다는 점이다. **구글의 진짜 노림수는 메타와 같은 빅테크 라이벌의 자본력을 담보로 최첨단 실리콘의 미래 공급량 전체를 선매수하는 공급망 선점에 있을 수 있다.**

이는 엔비디아가 GPU 시장에서 구사했던 진입장벽 구축 전략과 판박이다. 구글은 3단계 전략을 통해 TSMC 웨이퍼 시장의 게이트키퍼로 부상하고 있다.

1. 담보 확보: 메타, 애플과 장기 클라우드 계약을 체결하며 연간 20만 개 이상의 TPU 의무 구매 조항을 확보. 이는 100억 달러 이상의 확정 수요를 의미.

2. 협상력 확보: 이 확정된 수요를 무기로 TSMC에게 N2(2나노) 생산 라인 용량의 25%를 요구. TSMC 입장에서는 공장 가동률 보장과 투자 리스크 제거라는 확실한 이점이 있기에 이를 거부할 이유가 없음.

3. 경쟁자 차단: 결과적으로 Groq, Cerebras 등 AI 칩 스타트업들은 TSMC로부터 "구글이 향후 24개월 물량을 선점했다"는 답변을 듣게 될 가능성.

아무리 혁신적인 아키텍처를 설계해도 웨이퍼를 구할 수 없다면 무용지물이다. "가격보다 중요한 것은 접근성"이라는 원칙이 웨이퍼 시장에도 적용되면서, 스타트업들은 구조적인 '웨이퍼 절벽'에 직면하게 될 수 있다는 말이다.

(3) 전력 제약 시대의 셈법: 엔비디아가 진짜 강한 이유

전기가 아껴야 할 공과금이라기 보다, 가장 희소한 생산 자원이 된 시대에 칩의 효율은 단순한 스펙이 아니라 자본 생산성 지표(ROIC)가 된다. "동일한 전력 한도 내에서 얼마나 많은 지능(토큰)을 생산할 수 있는가"가 데이터센터의 매출 상한선을 결정짓기 때문이다.

이는 젠슨 황의 "토큰당/와트당" 철학으로도 설명된다. 이러한 **전장에서 아직 엔비디아의 우위는 여전히 견고하다. 핵심은 데이터센터의 물리적 한계인 '전력 제약'과 '토큰 수익(Token Revenue)'의 상관관계임을 잊어서는 안 된다**.

시스템이 거대해지고 복잡해질수록, 단순히 칩 하나(ASIC)를 잘 만드는 것만으로는 경쟁이 불가능해진다. 제한된 랙 공간에 어떤 칩을 채워 넣어야 '순수익(Net Profit)'이 극대화되는지가 칩 구매 의사결정의 핵심이 된다.

극단적인 디플레이션 시나리오를 가정한 시뮬레이션 결과는 충격적이다. 초기 비용 400만 달러인 엔비디아 NVL72 랙과, 공짜(0달러)로 제공되는 경쟁사 칩을 비교했을 때, 엔비디아를 선택하는 쪽이 오히려 낫다는 결과다. NVL72 랙을 사다 쓸 경우 3년간 900만 달러(약 120억 원)의 추가 이익을 창출하는 것으로 나타났다.

표 3. 전력 제한 하에서의 칩 수익성 비교: 공짜 칩 vs 엔비디아 NVL72

구분	엔비디아 NVL72	경쟁사 칩 (성능 50%)
초기 비용 (Capex)	4M (유료 구매)	0 (무료 제공 가정)
처리 성능	150만 토큰/초 (100%)	75만 토큰/초 (50%)
3년 누적 매출	약 26M	약 13M (성능 비례)
3년 순수익 (Net Profit)	22M	13M
결과 (차액)	**엔비디아 시스템을 쓰는 게 9백만 달러 우위**	

자료: 미래에셋증권 리서치센터

공간과 전력이 100MW로 고정된 상황에서 성능이 50%에 불과한 칩을 사용하는 것은 전기료 절감이 아니라, 벌 수 있었던 막대한 매출 기회비용(Opportunity Cost)을 날리는 행위이기 때문이다.

"공짜로 줘도 GPU 사서 쓰는 게 좋습니다"라고 주장한 젠슨 황의 수리적 근거다. 그래서 우리는 경쟁 칩(TPU 등)의 성능 향상과 토큰 가격(토큰 당 매출) 하락 속도에 따른 '경제적 해자 매트릭스'를 분석해봤다.

표 4. 엔비디아의 경제적 해자 매트릭스: 가격 하락과 경쟁 칩 성능의 함수

	GPU 대비 경쟁칩(XPU)들의 성능					
연간 토큰 생성 비용 하락률	구분	50%	60%	70%	80%	90%
	70%	9.15M	6.52M	3.89M	1.26M	-1.37M
	80%	7.73M	5.39M	3.04M	0.69M	-1.65M
	85%	7.09M	4.87M	2.66M	0.44M	-1.78M
	90%	6.50M	4.40M	2.30M	0.20M	-1.91M
	95%	5.96M	3.97M	1.97M	-0.02M	-2.01M
	99%	5.56M	3.65M	1.73M	-0.18M	-2.09M

자료: 미래에셋증권 리서치센터

- 안전 지대: 경쟁 칩 성능이 엔비디아의 70% 미만인 경우, 토큰 가격이 매년 99%씩 폭락하는 대재앙이 와도 엔비디아를 쓰는 게 이득.

- 붕괴 지대: 경쟁 칩 성능이 엔비디아의 90%에 도달하면, 토큰 가격이 가장 완만하게 하락한다고 가정해도(매년 70% 하락)에서도 엔비디아를 쓰는 게 손해. 이때 고객들은 GPU를 버리고 값싼 경쟁 칩을 선택하는 것이 합리적.

그러나 **엔비디아가 Blackwell과 그 후속작 Rubin을 통해 경쟁사(TPU 등)와의 성능 격차(토큰당/달러당/와트당 칩 효율)가 계속 20~30% 이상 유지한다면, 전력 제약이 심화되는 미래에도 엔비디아의 '공간당 수익 독점'은 수학적으로 깨지지 않는 철옹성이 될 가능성이라 할 수 있다.**

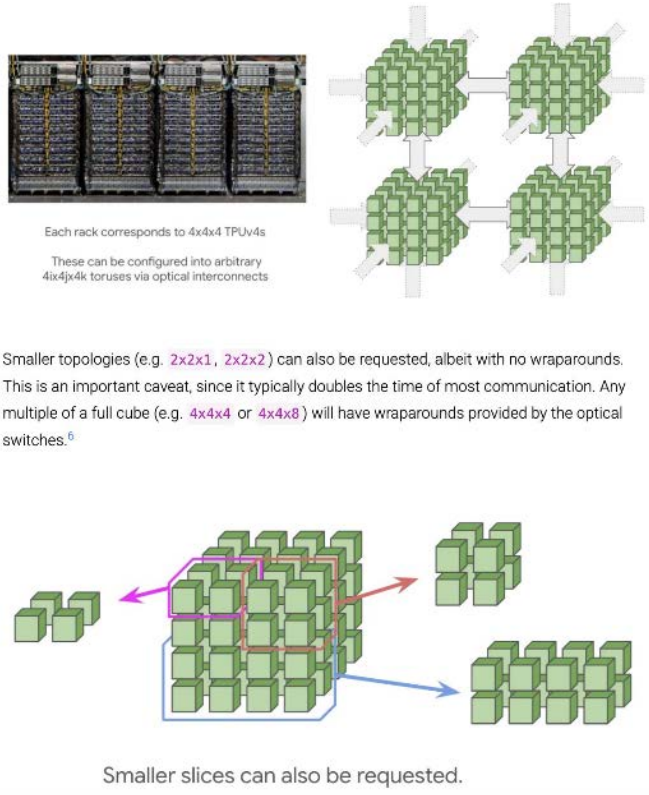
(4) TPU의 딜레마와 시장의 재편: 붕괴가 아닌 분할

TPU v7이 강력하지만, 아킬레스건은 존재한다. 바로 **최신 AI 추론 트렌드인 초저정밀도(FP4/FP6) 연산을 지원하지 않는다는 점이다.** 지난 번 우리가 AI Bi-weekly에서 강조했듯, Moonshot(Kimi)과 같은 최첨단 스타트업들은 이미 FP4로 서비스를 구동하고 있다. 앞으로 더욱 많은 기업들이 FP4 사용을 지향할 것을 보이기 때문에, 추론 시장을 엔비디아 GPU에게 생각보다 많이 못 뺏어올 전망도 가능하다.

구글의 TPU는 Lumentum의 OCS(광 회로 스위치) 기술을 활용해, 마치 3차원 도넛(Torus) 모양의 그물망처럼 9천 개 이상의 칩을 직접 연결하는 '슈퍼포드(SuperPod)' 구조를 만든다. **중앙 스위치를 거치지 않고 이웃된 칩끼리 직접 통신하는 식이다**.

이는 데이터 전송 속도를 극한으로 끌어올려 거대 모델 학습에 유리하지만, 동시에 외부 개발자들에게는 진입 장벽이 된다. **일반적인 '스파인-리프(Spine-Leaf)' 네트워크 구조와 달라 프로그래밍 난이도가 매우 높고, 유연성이 떨어져 다루기 까다롭기 때문이다**.

그림 25. 구글의 3차원 Torus 토폴로지와 광학 스위치(OCS)의 활용
수천 개의 칩을 4x4x4 큐브 단위로 묶고, 이를 광학 스위치(OCS)로 연결하여 'Superpod'를 구성
OCS를 통해(빛의 경로 변경) 칩 간의 물리적 연결(Topology)을 소프트웨어적으로 자유자재 재구성
이로써 작은 작업부터 초대형 모델 학습까지 유연하게 대응할 수 있다



자료: 구글, 미래에셋증권 리서치센터

TPU와 같은 XPU(각 업체들의 전용 AI 가속기)가 엔비디아에게 실질적인 위협이 되려면 단순히 성능만 좋아서는 안 된다. '규모의 경제'를 달성해야 한다. 전 세계적으로 4백만 명이 넘는 개발자들은 이미 엔비디아의 소프트웨어 생태계인 CUDA에 익숙해져 있다. 이들이 하루아침에 생소한 XPU 기반 프로그래밍으로 갈아타는 것은 불가능에 가깝다. 엔비디아 GPU는 여전히 가장 쉽고 범용적으로 AI를 개발할 수 있는 '표준 도구'의 지위를 굳건히 지키고 있다.

따라서 **일각에서 제기되는 "TPU가 엔비디아 제국을 무너뜨릴 것"이라는 주장은 과장된 측면이 있다**. 시장은 한쪽이 망하고 다른 쪽이 독식하는 것이 아니라, '역할 분담'을 통해 명확히 나뉠 것으로 사료된다. 엔비디아는 누구나 쉽게 쓸 수 있는 범용성과 최첨단 추론 기술을 앞세워 클라우드 시장을 장악할 것이다.

반면 구글 TPU는 내부의 거대 모델 학습과 대규모 추론 서비스를 도맡을 것이다. 또한 **구글은 물론 메타나 Anthropic과 같이 TPU를 보유하고 있는 쪽에서는 엔비디아와 GPU 가격을 협상할 때 강력한 카드로 활용될 것으로 예상한다.**

5. 승자의 조건: 결국에는 '메모리'와 '광통신'으로 흐른다

(1) 제로섬 게임이 아닌 '인프라 총량'의 폭발

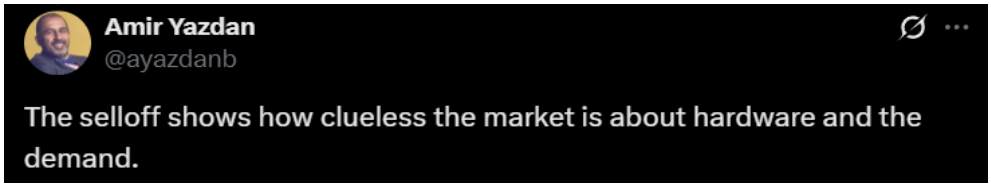
젠슨 황의 "우리는 고객사의 내부 팀(구글 TPU, AWS Trainium)과 경쟁한다"는 발언은 독점력이 영원하지 않으며, 시장이 범용 GPU와 커스텀 ASIC으로 양분되고 있음을 어느정도 인정한 것으로 볼 수 있다.

그림 26. Gemini 3.0 Pro 발표 직후, 마이크로소프트, 엔비디아, Anthropic 간 전략적 파트너십
이 세 기업 모두 구글의 풀스택 역량을 상당히 경계한다는 방증일 수 있음
Anthropic이 Azure 클라우드 컴퓨팅 용량 300억 달러 구매 약정 및 최대 1GW 규모 확장 계획
Anthropic과 엔비디아는 최초로 기술 파트너십 체결, AI 모델 최적화 및 차세대 하드웨어 공동 개발
엔비디아 최대 100억 달러, Microsoft 최대 50억 달러를 Anthropic에 투자

자료: 사티야 나델라, 미래에셋증권 리서치센터

이때 투자자들은 구글(TPU)과 엔비디아(GPU) 중 승자를 골라야 한다는 압박을 느끼지만, 인프라 관점에서 이 경쟁은 제로섬이 아니다. 오히려 '인프라 총량 및 TAM(Total Addressable Market)의 폭발적 증가'를 의미한다고 생각한다.

그림 27. 엔비디아 주가 하락에 관한 "구글의 TPU 담당" 임원의 트윗
"이번의 (엔비디아에 대한) 매도세는 하드웨어와 (컴퓨팅) 수요에 대해
시장이 얼마나 무지한지를 보여줍니다."



자료: Amir Yazdan(TPU를 설계하는 구글 딥마인드의 연구 과학자), 미래에셋증권 리서치센터

지금껏 '연산 장치(Compute)' 중심의 1차전이었다면, 2025년 말부터는 본격적으로 '대역폭의 전쟁(War of Bandwidth)' 국면이다. 구글의 TPU 포드든, 엔비디아의 Rubin 클러스터든, 거대한 연산 괴물들이 제 성능을 내기 위해서는 데이터가 이동하는 '길(Lane)'이 필수적이다.

표 5. Amin Vahdat(구글 AI 인프라 부사장)이 발표한 내용과 구글의 실적을 교차검증하여 정리

지표 (Metric)	수치 (Figure)	상세 내용 (Details)
목표 확장 속도	**6개월마다 2배**	**현재 서빙(Serving) 용량을 6개월마다 2배로 늘려야 수요 충족 가능**
중장기 목표	**1,000배**	**향후 4~5년 내 컴퓨팅 파워를 현재의 1,000배로 확장**
2025년 CapEx	92B	2025년 연간 총 자본지출 예상치. 기존 85B에서 대폭 상향 조정됨
Q3 2025 CapEx	24B	2025년 3분기에만 24B 집행. 이 중 60%가 서버/칩, 40%가 데이터센터/네트워크
클라우드 매출	15.16B (Q3 2025)	전년 대비 급증, 시장 예상치(14.7B) 상회. Gemini API 수요가 견인.
차세대 칩	TPU v7 'Ironwood'	2018년 TPU 대비 전력 효율 30배 개선. 본격 배치 시작.

자료: 구글, 미래에셋증권 리서치센터

따라서 투자자로서 편안한 포지션은 "누가 이겨도 돈을 버는" 영역이다. 우리는 그 영역에서 **커다란 파이를 차지할 곳이 메모리와 광통신이라 생각한다**.

(2) 메모리 반도체: 전략적 필수재로 상급지 이동

AI 모델이 고도화된다는 것은, Gemini 3.0 Pro에서 보듯이 모델의 뇌 용량(파라미터)이 비대해지고, 한 번에 기억하며 처리해야 할 정보량(Context Window)이 기하급수적으로 늘어남을 의미한다. 이로 인해 메모리 반도체는 이제 부품(Commodity)이 아닌, 인프라의 핵심이자 대체 불가능한 '전략 자산'으로 그 위상이 격상되었다.

구글이나 아마존 같은 클라우드 거인들에게 메모리 확보는 더 이상 비용 절감의 대상이 아니라 사업의 존폐를 가르는 필수 생존 조건이 된 것으로 풀이된다. 이미 현재 시장에서는 돈을 더 지불해도 물건을 구하기 힘든, 가격에 상관없이 확보해야 하는 거의 '완전 비탄력적' 공급 부족 국면에 진입했다.

이와 관련해, **구글 딥마인드의 데미스 하사비스 CEO는 "미래 AI 전쟁의 승패는 누가 더 정교한 '월드 모델'을 구축하느냐에 달려 있으며, 이를 뒷받침할 거대한 메모리 없이는 불가능하다"고 선언했**다. 월드 모델이란 AI가 머릿속에 현실 세계를 시뮬레이션하는 가상 공간이다. 기존 챗봇이 텍스트 확률 통계로 "사과를 놓으면 떨어진다"고 답했다면, 월드 모델 AI는 머릿속에서 사과를 떨어뜨려 보고 깨지는 장면까지 예측한 뒤 행동을 결정한다. 텍스트를 넘어 중력과 마찰력 같은 물리 법칙을 이해하는 '물리적 지능(Physical Intelligence)'으로 진화하려면, 3차원 공간과 시간 정보까지 모두 담을 수 있는 압도적인 메모리 용량과 대역폭이 필수적이라는 말이다.

이는 필연적으로 HBM(고대역폭 메모리) 수요의 폭발로 이어진다. 구글의 최신 **TPU v7은 엔비디아 GPU보다 HBM 탑재 비중이 높으며, 엔비디아 역시 2027년 출시될 Rubin Ultra에 무려 1TB의 HBM을 탑재할 예정이다**. 칩의 로고가 구글이든 엔비디아든 상관없이, 그 안에 들어가는 HBM과 데이터를 저장하는 eSSD의 수요는 상수가 아닌 지수함수로 증가할 가능성이 매우 높다.

특히 2026년 말 차세대 HBM4 생산이 시작되면 일반 D램 생산 라인까지 잠식하게 되어 범용 메모리 공급 부족은 더욱 심화될 것이다. 결국 **HBM은 AI 칩 하나당 차지하는 원가 비중이 급증하고 있고, 시스템 내에서 '가장 비싼 부동산'으로 자리 잡게 될 수도 있다**. 심지어 TSMC가 가져가는 몫보다 메모리 업체들이 가져가는 비중이 커질 것이라는 전망까지 나온다. 이는 메모리 3사의 협상력이 역사상 최고조로 진행되고 있음을 의미하며, AI 하드웨어 시장의 부가 로직 칩에서 메모리로 이동하는 '부의 재분배'가 시작되었음을 시사한다.

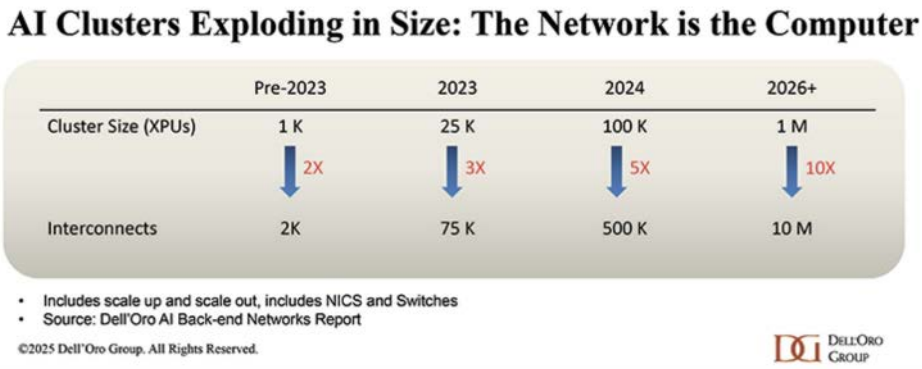
(3) 광학 솔루션: 빛의 시대

현재 AI 인프라 시장은 기존 800G 속도의 통신망이 성숙기에 접어듦과 동시에, 두 배 더 빠른 1.6T(테라비트) 속도의 차세대 통신망이 열리는 '슈퍼 커넥티비티' 구간에 진입했다. 여기서 발생하는 부가가치는 AI 칩 못지않게 강력하다.

반도체 칩의 성능은 비약적으로 향상되었지만, 데이터를 실어 나르는 구리선(Copper)의 속도는 이를 따라가지 못한다. 결국 **데이터센터 내부 연결망은 선택이 아닌 필수로 '광통신(Optics)', 즉 빛으로 데이터를 쏘는 방식으로 전환되고 있다**. 막대한 전력 소모 때문에 여러 건물에 흩어놓은 GPU들을 마치 하나처럼 작동시키려면, 이들을 초고속으로 묶어주는 광학 장비가 반드시 필요하기 때문이다. **통신 속도는 2025년 800G, 2027년 1.6T, 2029년 3.2T로 빠르게 세대교체가 이루어질 전망이다.**

특히 1.6T를 넘어 3.2T 속도로 가면 구리선으로는 신호를 10cm도 보내기 힘들어진다. 이는 매번 칩이 바뀔 때마다 구리선을 걷어내고 비싼 광모듈로 도배하거나, 아예 기판 레벨에서 광통신을 통합하는 CPO 기술 도입을 강제하게 된다. **GPU 클러스터 규모가 1,000배 커질 때 칩들을 연결하는 가닥 수는 5,000배가량 폭증**하므로, 칩을 사는 예산보다 칩들을 연결하는 네트워킹 예산이 훨씬 가파르게 증가하는 구조가 만들어질 수 있다.

그림 28. AI 클러스터 규모 폭발과 인터커넥트의 기하급수적 증가
2026년 AI 클러스터 규모(XPU 수)는 2023년 대비 40배 증가한 100만 개에 달할 것으로 예상
주목할 점은 칩의 개수보다 연결하는 인터커넥트의 수는 그보다 훨씬 가파른 속도로 폭증한다는 것

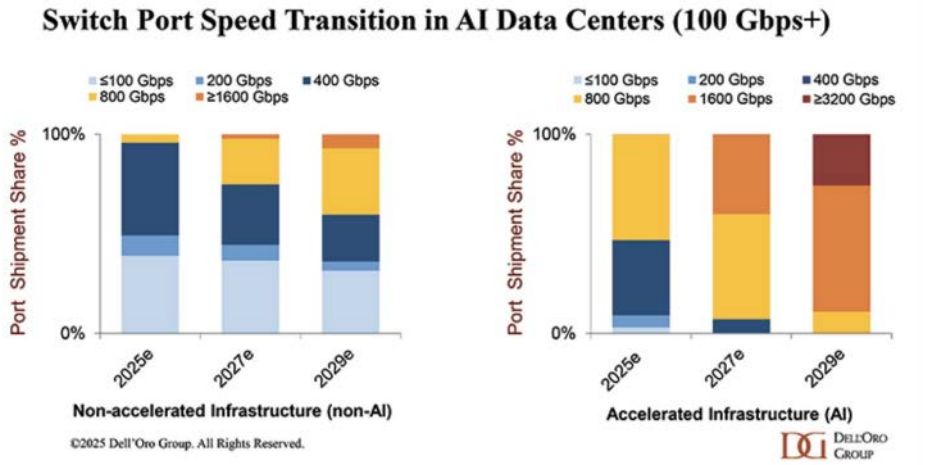


자료: Dell'Oro Group, 미래에셋증권 리서치센터

이러한 흐름 속에서 주목해야 할 기업들이 있다. **Lumentum(LITE)은 구글 TPU 인프라의 핵심인 OCS(광 스위치) 시장을 선도**하며, 엔비디아 생태계 밖에서도 수익을 내는 기업이다. 구글은 칩과 칩 사이를 전기가 아닌 빛으로 직접 연결하는 기술에 베팅했는데, 루멘텀은 이 독특한 기술의 수혜를 받게 된다. **동시에 이 회사는 엔비디아용 800G 모듈 레이저 공급망의 40%를 점유**하며 양쪽 진영 모두에서 이익을 취하는 양수겸장의 포지션을 구축했다.

그리고 Coherent(COHR)는 부품부터 모듈까지 수직 계열화된 생산 능력을 바탕으로, 엔비디아와 구글 양쪽의 폭발적인 800G/1.6T 모듈 수요를 감당할 수 있는 기업이다. 미중 갈등으로 인해 중국 경쟁사들의 리스크가 커질수록, 미국 기업인 Coherent의 전략적 가치는 더욱 상승할 가능성이 높다.

그림 29. AI 데이터센터 내 스위치 포트 속도의 세대교체
AI 가속 인프라 내 스위치 포트 속도는 비 AI 인프라와 달리 극적으로 빠른 전환
2025년에는 800Gbps가 주류, 2027년에는 1.6T가 과반을 차지, 2029년에는 3.2T시대가 열릴 것
초고속 통신 환경에서는 구리선의 물리적 한계로 인해 광통신(Optics) 도입이 필수적



자료: Dell'Oro Group, 미래에셋증권 리서치센터

빛으로 쏘아진 광신호를 다시 칩이 이해할 수 있는 전기신호로 변환하고, 이 과정에서 깨지고 뭉개진 신호를 수학적으로 완벽하게 복원하는 반도체가 바로 DSP(Digital Signal Processor) 칩이다. 데이터 전송 거리가 수십 킬로미터로 길어지고 속도가 800G 이상으로 빨라지면, 빛은 공기 중의 산란이나 광섬유의 굴절로 인해 '너덜너덜'해진다. 특히 빛은 파장(색깔)마다 속도가 미세하게 달라, 빨간색은 빨리 도착하고 파란색은 늦게 도착하며 신호가 뒤섞이는 '분산(Dispersion)' 현상이 발생한다.

DSP는 이때 개입하여 늦게 온 신호를 당겨주고 빨리 온 신호를 밀어주는 '분산 보상' 알고리즘을 수행한다. 또한, 신호가 너무 약해져 0인지 1인지 구분하기 어려울 때, DSP는 고도의 확률 통계를 돌려 "이건 문맥상 1일 확률이 99.9%다"라고 판단하고 원본 데이터를 복구하는 '오류 정정(FEC)' 기능을 수행한다.

초당 8,000억 개의 비트가 쏟아지는 극한의 상황에서 실시간 미분/적분을 통해 깨끗한 신호를 만들어내는 이 칩은, 전 세계에서 Marvell을 포함해 서너 개 회사만이 만들 수 있는 진입 장벽이 매우 높은 기술이다. **구글 TPU든 엔비디아 GPU든, 초고속으로 데이터를 주고받으려면 반드시 이 DSP를 거쳐야 한다는 게 포인트다**. 데이터가 이동할 때마다 칩을 통과해야 하므로, DSP 관련 기업들은 AI 트래픽이 늘어날수록 확실하게 돈을 버는 '통행료' 비즈니스 모델을 갖게 될 것으로 보인다.

그리고 **Broadcom(AVGO)은** 네트워킹의 장인이자, AI 맞춤형 칩(ASIC) 시장에서 TSMC와 같은 위상을 가진 기업이다. **구글 최신 TPU v7의 공동 개발자로서 핵심 설계와 물리적 구현을 담당한 사실이 확인되었으며, 애플, xAI, OpenAI 등 업계 거물들 모두가 브로드컴을 파트너로 선택했다.** 엔비디아를 제외한 모든 빅테크 기업들이 자체 AI 클러스터를 구축할 때 1순위로 채택하는 표준 플랫폼이며, 구글이 자체 칩을 늘릴수록 브로드컴의 매출 또한 비례해서 증가하는 구조를 갖고 있다.

- 브로드컴(AVGO)은 구글이 설계한 논리적 '엔진 설계도(Verilog 코드)'를 넘겨받아, 이를 실제 작동하는 물리적 반도체 칩으로 구현(Physical Design)하는 '마스터 빌더' 역할을 수행. 수억 개의 트랜지스터를 배치하고 연결하는 복잡한 공정을 전담하며, 칩의 전력 효율과 수율을 책임.

또한, TPU가 외부와 통신하는 데 필수적인 고속 인터페이스(SerDes, PCIe)와 같은 핵심 IP 블록을 제공하여 칩의 '뼈대'를 완성. 구글이 AI 연산 로직에만 집중할 수 있도록, 브로드컴은 그 외의 모든 물리적 난제와 표준 연결성을 해결해주는 파트너. F1 레이싱팀(구글)이 엔진을 설계하면, 전문 제작사(브로드컴)가 섀시와 구동계를 결합해 완벽한 경주차를 조립해내는 것과 유사.

표 6. 하이퍼스케일러와 브로드컴의 XPU 개발에 있어서의 분업

구분	빅테크 (Brain Designer)	브로드컴 (Master Builder)
핵심 역할	"무엇을, 어떻게 계산할 것인가?" (WHAT to compute)	"설계도를 가지고 어떻게 물리적 칩으로 만들고 연결할 것인가?" (HOW to build)
주요 업무	**AI 아키텍처 설계, Verilog(RTL) 코딩**	**Physical Design (Place & Route), IP 블록 제공 및 통합**
결과물	TPU의 성능과 효율을 결정하는 논리 설계도	실제 작동하는 물리적 칩과 그 안정성, 연결성

자료: 미래에셋증권 리서치센터

결론적으로 시장에서 안전한 초과 수익(Alpha) 전략은 챔피언이 누가 될지 맞히는 것이 아니라, 링 위에 오르는 모든 선수가 반드시 착용해야 할 글러브와 마우스피스를 파는 기업에 투자하는 것이지 않을까? **구글 TPU의 약진이 엔비디아에게는 당장의 점유율 위협일지 몰라도, 메모리와 광통신 기업들에게는 '또 하나의 거대한 고객'이 등장한 긍정 요소로 치환될 것으로 기대한다.**

표 7. TPU의 부상 속 관련주 소개

종목명 (티커)	AI 인프라에서의 역할
Broadcom (AVGO)	구글 TPU, 메타 MTIA, 애플/xAI 칩 설계 파트너. 엔비디아 점유율 하락 시 수혜. Arista, Cisco 스위치에 들어가는 칩(Tomahawk) 독점. 1.6T/3.2T 업그레이드 사이클의 수혜.
Micron (MU)	CEO가 '2026년 이후까지 공급 부족'이라고 선언. HBM용 웨이퍼를 선점해 레거시 DRAM 가격까지 동반 상승.
Lumentum (LITE)	TPU 클러스터를 연결하는 OCS(광 스위치)를 독점 공급. 칩 외부에서 레이저를 쏘는 ELS 시대, 고출력 InP 레이저 시장을 장악. 스위치당 BOM 비용 상승의 직접 수혜.
Celestica (CLS)	TPU 랙 조립 및 시스템 통합을 담당하는 핵심 파트너. 단순 조립을 넘어 AI 백엔드 스위치 시장 점유율 3위권 안착.
Vertiv (VRT)	전력 밀도 상승으로 액체 냉각이 필수 인프라가 됨. 칩 종류와 상관없이 모든 랙에 필요. 쿨링, 전력, 랙을 턴키 패키지로 공급.
Arista (ANET)	마이크로소프트·메타가 선호하는 개방형 이더넷 스위치 강자. 엔비디아 Spectrum-X의 공격적인 번들 전략으로 점유율 방어 경쟁 심화.
Coherent (COHR)	데이터센터 간 연결(DCI)에 쓰이는 장거리 광모듈의 대표 공급업체. Lumentum과 함께 ELS 시장에 진입. 제품 포트폴리오가 넓어 변동성은 상대적으로 낮음.
TSMC (TSM)	칩 제조뿐 아니라 광 패키징(실리콘 포토닉스)까지 수직 통합하며 파운드리 영향력 확대. 엔비디아든 구글이든, 대부분의 첨단 칩은 결국 TSMC에서 생산.
Marvell (MRVL)	장거리 전송 과정에서 깨지는 신호를 복구하는 DSP 칩 분야 글로벌 1위. 마이크로소프트 ASIC(Maia 300) 수주 실패 등 커스텀 칩 사업의 변동성 존재.
Ciena (CIEN)	데이터센터가 지리적으로 분산될수록(예: 마이크로소프트 Fairwater) 광 전송 장비 수요는 급증 전망.

자료: 미래에셋증권 리서치센터

II. AI Agents

1. 쇼핑 에이전트와 30분의 현실

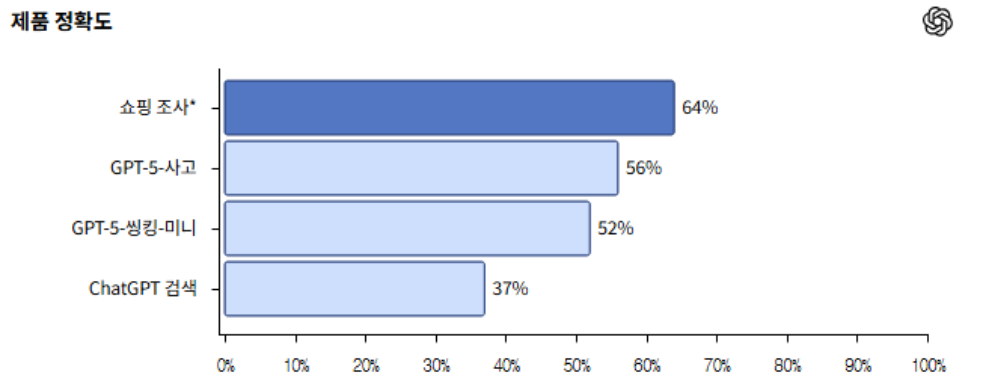
현재 생성형 AI 시장은 구글, 앤트로픽, xAI, OpenAI의 4강 과점 체제로 굳어졌으나, OpenAI의 상황은 녹록지 않다. 최근 모델 품질 평가에서 구글의 Gemini 3와 xAI의 Grok 4.1에 밀려 충격적인 3위를 기록했고, 야심차게 준비한 GPT-5.1 역시 시장의 판도를 뒤집기에는 역부족이라는 평가를 받았다. 구글은 '풀스택'을 완성했고, 앤트로픽은 '기업 고객'을 사로잡았으며, xAI는 '광속'으로 추격해오고 있다. 바야흐로 최고 지능을 향한 춘추전국시대, 왕좌를 위협받는 OpenAI의 반격 카드는 과연 어디를 향하고 있을까?"

이 질문에 대한 **OpenAI의 대답은 명확하다**. 그들은 모델 성능 경쟁이라는 진흙탕 싸움에서 **한발 물러나, 에이전트를 통한 B2C 영역의 완전한 확장으로 전선의 축을 옮겼다**. 위기 속에서 OpenAI가 선택한 **돌파구는 쇼핑 리서치라는 이름의 생활 밀착형 에이전트다.**

질문에 답하는 챗봇을 넘어, 사용자의 일상과 지갑을 여는 소비의 패러다임을 바꾸는 주역이 되겠다는 의도가 담겼다. GPT-5 Thinking mini에 기반한 이 에이전트는 사용자의 과거 대화와 기억을 바탕으로 맥락을 파악하고, 이를 기반으로 작은 아파트에 어울리는 조용한 청소기를 찾거나, 예술을 좋아하는 조카를 위한 선물을 고르는 등의 복잡하고 개인화된 요구를 처리한다.

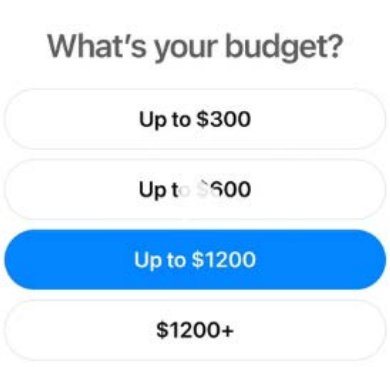
사용자의 요구사항을 충족하는 제품의 비율을 기준으로 제품 정확도를 측정했을 때, 일반적인 ChatGPT 검색 결과보다 약 70% 개선된 64%의 정확도를 보였다. 인터넷 검색 결과를 기반으로 가격, 재고, 리뷰를 분석하고, 사용자의 피드백에 따라 결과를 정교하게 다듬는다. ChatGPT Pulse 기능을 통해 사용자가 전기 자전거를 구매했다면 추가로 필요한 헬멧이나 액세서리를 선제적으로 제안하기도 한다.

그림 30. 쇼핑 리서치는 GPT-5-Thinking-mini에서 후처리된 미니 모델.
응답에서 사용자 요구사항(예: 가격, 색상, 재료, 사양 등)을 충족하는 제품 비율로 제품 정확도 측정



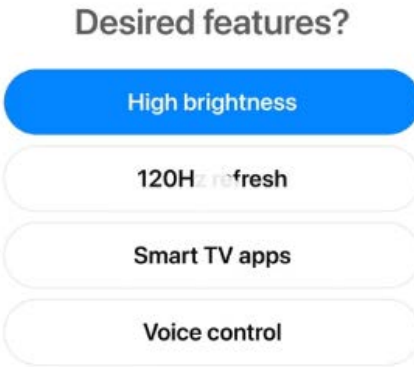
자료: OpenAI, 미래에셋증권 리서치센터

그림 31. 예산, 사용 용도 등을 질문



자료: OpenAI, 미래에셋증권 리서치센터

그림 32. 사용자의 답변을 확인하고,



자료: OpenAI, 미래에셋증권 리서치센터

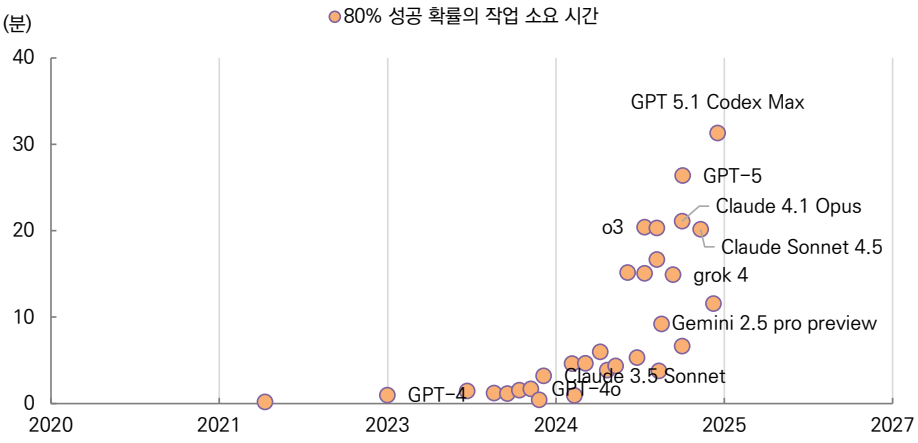
그림 33. 그에 부합하는 제품을 추천

자료: OpenAI, 미래에셋증권 리서치센터

그러나 동시에 OpenAI의 쇼핑 에이전트는 현 시점 AI 에이전트의 현주소를 가장 적나라하게 보여주는 사례다. 이 서비스는 분명 훌륭하다. 겨울용 텐트를 찾아달라는 모호한 요청에도 예산과 브랜드를 고려해 웹을 누비고, 몇 분 만에 꽤 그럴듯한 결과를 내놓는다. 전반적으로, **인간이 수행했을 때 약 30분 정도 걸릴 법한 검색과 비교 과정을 5~10분 남짓한 시간에 해결하고 간결하게 시각화해 보여준다**.

그러나 바로 이 30분이라는 시간은 현재 에이전트의 능력을 증명하는 동시에, 그 한계를 보여주는 숫자다. METR의 최신 연구가 지적했듯, 최신 AI 모델들은 몇 분에서 30분 내외의 과제에서는 인간을 능가하지만, 그 이상의 호흡이 필요한 과제에서는 급격히 무력해진다. 쇼핑 에이전트 역시 이 **30분의 벽 안에서는 유능하지만, 그 벽을 넘어서야 하는 순간 우리는 AI의 맹점을 발견하게 된다.**

그림 34. AI 모델이 80% 확률로 성공할 수 있는 작업을 인간이 처리하는 데 걸리는 시간. 이달 출시된 GPT 5.1 Codex Max조차 아직 31분 수준.



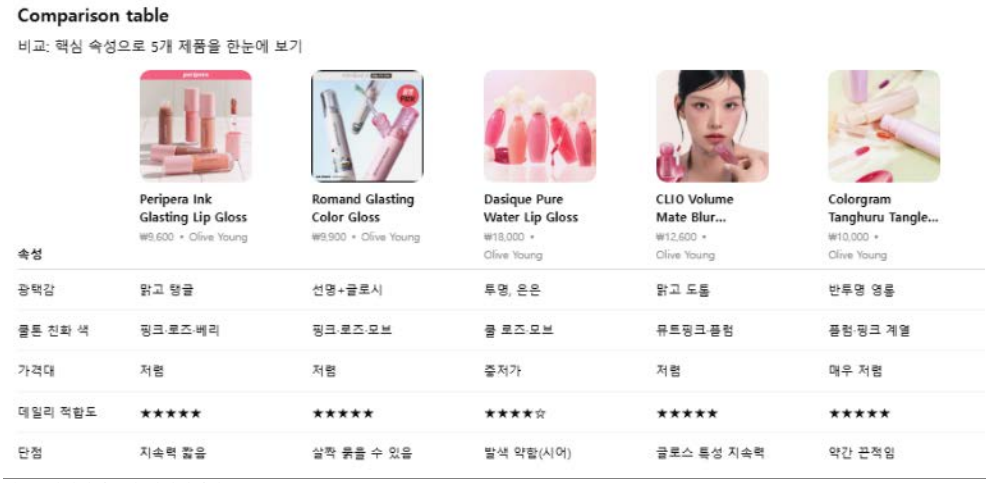
자료: METR, 미래에셋증권 리서치센터

2. 기억의 부재와 만족의 상대성

B2C 서비스에 있어, 우리는 이러한 에이전트의 한계를 만족의 상대성이라는 관점에서 바라볼 필요가 있다. OpenAI의 쇼핑 에이전트가 내놓은 결과물은 모든 사용자에게 동일한 만족을 주지 못한다. 그 **결정적인 변수는 사용자가 해당 분야에 대해 얼마나 깊이 알고 있느냐다**.

예를 들어 화장품에 정통한 일명 코덕(코스메틱 덕후)이 에이전트에게 수분 크림 추천을 의뢰했다고 가정해 보자. 에이전트가 30분 수준의 웹 검색으로 가져온 베스트셀러 리스트는 이들에게 아무런 감흥을 주지 못한다. 그들은 자신의 피부 타입, 과거에 트러블이 났던 성분, 선호하는 텍스처, 현재 사용하는 앰플과의 궁합까지 고려된 깊이 있는 제안을 원하기 때문이다. 반면 같은 사람에게 잘 모르는 로봇청소기를 추천해 달라고 하면, 에이전트가 가져온 일반적인 스펙 비교와 리뷰 요약만으로도 충분히 만족할 수 있다.

그림 35. 쇼핑 에이전트가 추천해 준 '쿨톤 틴트'들
대표적인 제품들을 잘 선정했지만, 구체적인 색상 추천이 빠져 있어 결국 사람이 다시 찾아봐야 한다는 점에서 디테일의 한계를 엿볼 수 있음



자료: 미래에셋증권 리서치센터

이 차이는 기억의 구조적 한계에서 기인한다. 물론 OpenAI의 쇼핑 에이전트는 과거 대화 기록을 참조한다고 주장한다. 하지만 여기에는 두 가지 맹점이 있다. 첫째는 현실적인 데이터의 부재다. 우리는 그동안 ChatGPT와 대화하며 내 피부 타입이 무엇인지, 내가 어떤 향을 싫어하는지 사소한 선호도를 일일이 입력하지 않았다. 따라서 에이전트 입장에서 현재의 사용자는 사실상 백지 상태에 가깝다.

둘째는 검색의 비효율성이다. 설령 과거에 그런 대화를 나눴다 하더라도, 수많은 대화 로그 속에 파편화된 정보를 매번 처음부터 다시 검색하는 것은 매우 비효율적이다. 이는 마치 도서관에서 책을 찾을 때마다 모든 책의 본문을 처음부터 끝까지 다시 읽는 것과 같다. 이중 첫 번째 문제는 서비스가 출시되고 시간이 흐를수록 나아질 수 있다. 원하는 상품과 모델이 찾아온 결과물에 대한 피드백이 쌓이면서 각 사용자별 컨텍스트가 생성되기 때문이다.

진짜 문제는 정보의 구조다. 결국 **에이전트가 코덕을(그리고 다른 사람들을) 만족시키려면, 산재된 대화 로그를 검색하는 것을 넘어, 사용자의 선호도를 별도의 구조화된 장기 기억 프로파일로 추출하여 저장하고 있어야 한다**. 현재의 에이전트는 아무리 똑똑해도 내가 지난달에 샀던 크림이 왜 별로였는지는 잊어버리기 쉬운 30분짜리 단기 아르바이트생에 불과하다. 더 긴 시간 동안 고민하고, 더 소비자에게 딱 맞는 결과를 가져오기 위해서는 단순한 로그의 축적을 넘어선 기억의 아키텍처가 필수적이다.

3. 처방 1: 조직적 기억, 개인적 의도

에이전트가 30분의 벽을 넘고 실무에 투입되려면, 가장 먼저 해결해야 할 것은 맥락을 장악하는 것이다. 지난 2주간 빅테크들의 행보는 사용자가 누구이며, 지금 무엇을 원하는가를 집요하게 파고드는 데 집중되었다. 흥미로운 점은 **마이크로소프트와 OpenAI가 각각 B2B와 B2C라는 서로 다른 전장에서 맥락을 해석하는 방식을 보여주었다**는 것이다.

(1) B2B의 기억: 마이크로소프트 Work IQ와 조직적 기억

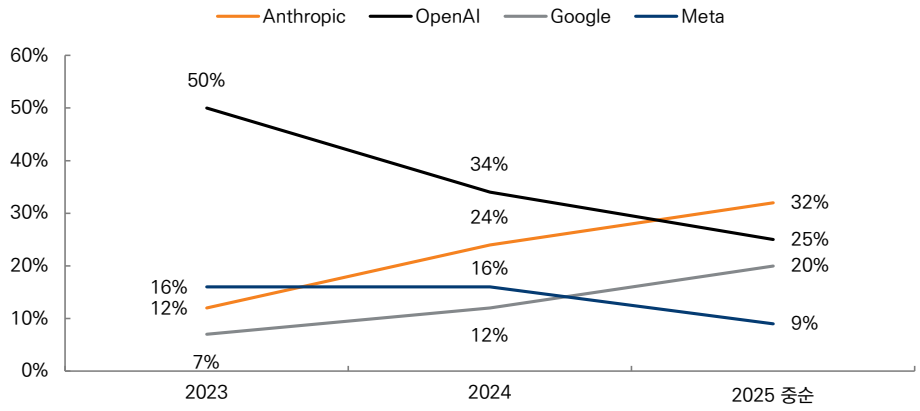
11월 18일 발표된 **마이크로소프트의 Work IQ는 기업용 에이전트가 나아가야 할 방향, 즉 조직적 기억의 구축을 명확히 제시한다**. 기업 환경에서 맥락은 단일한 파일이나 대화 속에 존재하지 않는다. 이메일, 회의록, 채팅, CRM 데이터 등 산재된 정보들 사이의 관계 속에 존재한다. Work IQ의 핵심은 바로 이 관계를 이해하는 것이다. 이는 Microsoft 365 내의 모든 데이터를 연결하는 거대한 신경망인 Microsoft Graph를 기반으로 작동한다. 여기에는 사용자, 그룹, 파일, 메시지, 일정, 할 일 등 업무와 관련된 수십억 개의 데이터 노드가 위치해 있다. 그리고 Microsoft Graph는 이 노드를 연결함으로써 노드 간의 관계와 맥락을 일종의 지도처럼 데이터화할 수 있다. 즉, Work IQ는 Microsoft식 '온톨로지' 구축이다.

여기에 그래프 그라운딩 기술을 적용시켜, 에이전트는 사용자의 모호한 지시를 조직의 데이터 맵 위에 정확히 위치시킨다. 예를 들어 사용자가 김 대리와 논의 중인 예산안을 물으면, 시스템은 단순히 예산안이라는 키워드가 포함된 파일을 찾는 것이 아니라, 김 대리라는 인물 노드와 최근 주고받은 채팅, 그리고 그 맥락에서 공유된 엑셀 파일 노드를 추적해 정확한 문서를 찾아낼 수 있다. 이 과정을 통해 에이전트는 단순한 **검색 도구를 넘어, 사용자의 업무 히스토리와 인적 네트워크를 구조적으로 기억하고 공유하는 동료로 거듭난다.**

(2) B2C의 의도: OpenAI 쇼핑 에이전트와 개인적 맥락

반면 **11월 24일 공개된 OpenAI의 쇼핑 리서치는 B2C 영역에서 개인의 의도를 파고드는 데 집중한 전략이다**. OpenAI가 이 시점에 쇼핑 에이전트를 꺼내 든 배경에는 절박한 시장 상황이 있다. 최근 데이터에 따르면 기업 내 LLM 점유율에서 OpenAI는 하락세인 반면 경쟁자들은 점유율을 늘리며 맹추격 중이다. B2B 시장에서의 독점적 지위가 흔들리는 상황에서, OpenAI는 가장 대중적이고 상업적인 쇼핑 영역을 통해 B2C에서의 주도권을 가져와야만 했다.

그림 36. 엔터프라이즈 LLM API 점유율 변화. OpenAI의 지분을 구글과 앤트로픽이 나눠 가진 상황



자료: Menlo Ventures, 미래에셋증권 리서치센터

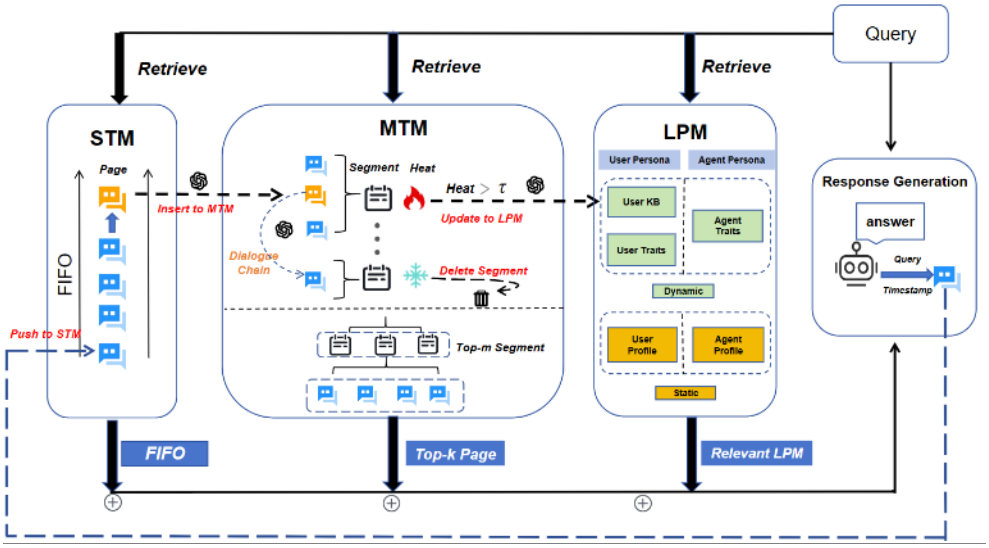
이 에이전트의 특징은 자율적인 다단계 추론에 있다. 겨울 캠핑용 텐트라는 모호한 요청이 들어오면, 에이전트는 이를 즉시 검색어로 변환하지 않는다. 대신 내부적으로 겨울이라는 계절적 특성, 사용자의 과거 선호 브랜드, 예산 범위라는 하위 과제로 분해한 뒤, 스스로 복수의 웹사이트를 탐색하고 장바구니에 담으며 정보를 비교 분석한다. 이는 사용자의 의도를 파악하고 실행 계획을 수립하는 에이전트 고유의 능력을 극대화하여 B2C 시장을 장악하겠다는 전략이다.

(3) 학계의 검증: 계층형 메모리 아키텍처

산업계가 B2B와 B2C에서 각각 기억의 문제를 풀려 노력하는 동안, 학계에서는 이를 뒷받침할 이론적 토대를 마련했다. EMNLP 2025에서 발표된 MemoryOS는 인간의 뇌 구조를 모방한 계층형 메모리 아키텍처를 통해 LLM의 성능을 비약적으로 높일 수 있음을 증명했다. 이 아키텍처는 **단기 작업 기억, 중기 에피소드 기억, 장기 의미 기억의 세 가지 레이어로 구성되며, 각 레이어 간의 동적 정보 이동을 통해 맥락을 유지한다**.

화장품 추천을 예로 들면, 에이전트가 수만 건의 대화 로그를 매번 뒤지는 것이 아니라, 장기 기억 슬롯에서 사용자의 민감성 피부 타입을 즉시 꺼내오고 중기 기억에서 최근 선호하는 브랜드 경향을 파악하는 식이다. 이는 **우리가 지난 보고서에서 하드웨어적 접근으로 제안했던 구조와 정확히 맞닿아 있으며,** 마이크로소프트와 OpenAI가 추구하는 기억하는 에이전트를 구현하기 위한 필수적인 설계도라 할 수 있다.

그림 37. 장기 기억을 위한 계층형 메모리 아키텍처. 단기 기억에서 추상화한 정보를 중기/장기 기억으로 저장



자료: 'Memory OS of AI Agent', 미래에셋증권 리서치센터

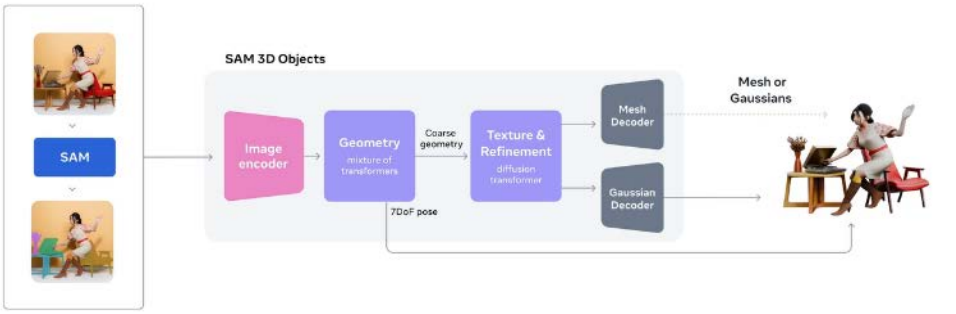
4. 처방 2: 세상을 보는 눈과 뇌

기억 다음으로 필요한 것은 감각이다. 지난 보고서에서 우리는 텍스트 데이터는 정적이기 때문에 시간의 흐름에 따른 변화와 인과관계를 학습하기 어렵다고 지적했다. 텍스트로 학습한 AI는 유리잔이 떨어지면 깨진다는 명제적 사실은 알지만, 바닥의 재질과 높이에 따라 파편이 어떻게 튀는지 예측하여 발을 피하는 물리적 상식은 결핍되어 있다. 지난 Bi-Weekly에서 이를 극복하기 위해 영상 모달리티로의 확장이 필요하다고 주장했었다. 이번 주 발표된 기술들은 이 주장을 현실로 만들고 있다.

(1) SAM 3과 SIMA 2: 3차원의 생성과 이해

11월 19일 메타가 공개한 SAM 3D는 AI의 시각 능력을 기존의 평면적 인식에서 입체적 공간지각으로 격상시켰다. 이전 버전인 SAM 2가 비디오 내 객체의 픽셀을 추적하는 데 그쳤다면, SAM 3D는 **단 한 장의 2D 이미지나 비디오 입력만으로 대상의 3D 포인트 클라우드와 메쉬를 실시간으로 생성**해 낼 수 있게 됐다. 이 모델의 핵심 아키텍처는 2D 이미지 인코더와 3D 기하학 디코더의 결합이다. 2D 인코더가 이미지 내의 시각적 특징을 추출하면, 3D 디코더가 이를 깊이 정보와 결합하여 입체적인 구조물로 재구성한다.

그림 38. 들어온 2차원 이미지를 기반으로 3D 기하학 디코더가 깊이 정보와 결합해 입체로 재구성



자료: Meta, 미래에셋증권 리서치센터

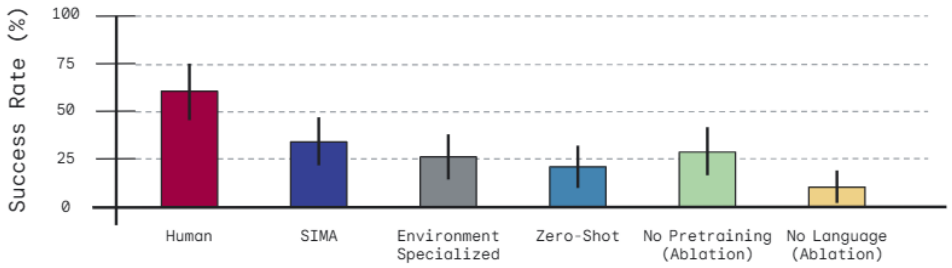
더 나아가 개념 프롬프트 기술이 적용되어, 사용자가 '작은 나사'나 '태양광 패널' 같은 텍스트만 입력해도 화면 내의 해당 객체를 찾아내 3D 형상으로 분할한다. **이러한 3차원 재구성 능력은 물리 세계에서의 AI 성능을 개선시킬 수 있다**. AI가 탑재된 로봇이 세상을 바라볼 때 색깔 정보의 나열인 픽셀이 아니라, 부피와 표면과 깊이를 가진 잡을 수 있는 객체로 인식할 수 있기 때문이다. 로봇 팔이 컵의 손잡이 위치를 정확히 파악하고 파지하는 조작 행위의 시각적 토대가 완성된 것이다.

또한, 2차원 이미지를 기반으로 3차원 에셋을 재구성할 수 있기 때문에, 부족했던 3차원 공간 데이터를 대량으로 찍어낼 수 있다는 점 역시 성능 개선에 일조할 것으로 기대된다.

시각 정보가 확보되었다면 이를 해석할 뇌가 필요하다. 구글 딥마인드의 SIMA 2는 시각 정보를 바탕으로 물리적 세계를 이해하는 에이전트를 공개했다. SIMA 2는 시각-언어-행동 모델을 기반으로 하며, 게임과 같은 가상 세계에서의 무수한 시행착오를 통해 학습한다.

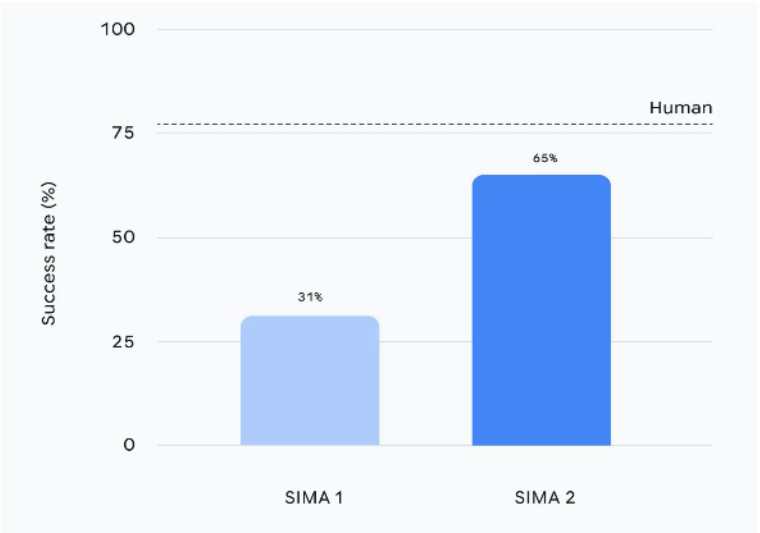
특히 주목해야 할 점은 자연어로 주어진 명령에 대해 **AI 에이전트가 스스로 목표를 설정하고 행동한 뒤, 그 결과가 물리 법칙에 부합하는지 혹은 목표를 달성했는지에 따라 스스로 보상을 부여**한다는 점이다. 다양한 생존 게임에서 모델이 임무를 얼마나 잘 수행하는지를 평가한 결과, SIMA 1에서는 모델 성능이 인간의 절반 수준에 불과했지만, 이번에 공개된 SIMA 2는 **65%의 업무 성공률을 보여 인간 수준에 한층 더 가까워진 모습을 보여줬다**.

그림 39. SF 생존 게임인 'No Man's Sky'에서의 임무 성공률
SIMA 1은 인간 수행 능력의 절반 수준에 불과했음



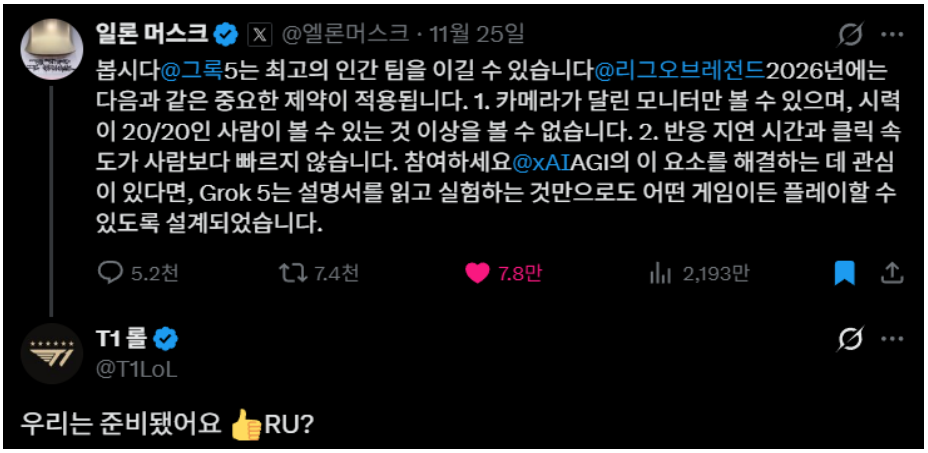
자료: Scaling Instructable Agents Across Many Simulated Worlds, 미래에셋증권 리서치센터

그림 40. 그러나 SIMA 2에서는 업무 수행 능력이 65%까지 상승하며, 인간 수준에 근접



자료: Google, 미래에셋증권 리서치센터

그림 41. 리그오브레전드(LoL) 챔피언 'T1' 팀과 맞붙고 싶다고 도전장 내민 xAI의 일론 머스크



자료: 일론 머스크, T1, 미래에셋증권 리서치센터

5. 결론: 기억을 선점하는 자가 시대를 지배한다

지난 보고서에서 우리는 **해마가 없는 AI, 상식이 없는 AI는 반쪽짜리**라고 진단했다. 그리고 불과 2주 만에, 시장은 마이크로소프트의 Work IQ, OpenAI의 쇼핑 에이전트, Meta의 SAM 3과 딥마인드의 SIMA 2를 내놓으며 우리의 생각과 궤를 같이하고 있다. 이는 거대 언어 모델의 파라미터 경쟁이 한계 효용에 도달했으며, 이제는 그 지능을 담을 그릇인 시스템 아키텍처 경쟁으로 전장이 이동했음을 의미한다.

에이전트가 가야 할 로드맵은 더욱 선명해졌다. 첫째는 기억이다. 우리가 앞선 예시에서 확인했듯, 사용자에게 진정한 효용을 주는 것은 30분의 현란한 검색 능력이 아니라 축적된 맥락에 대한 이해다. OpenAI가 쇼핑 에이전트를 통해 개인의 취향을 구조화하려 하고, 마이크로소프트가 Work IQ를 통해 기업의 업무 맥락을 연결하려 하는 것은 결국 누가 먼저 사용자의 해마를 선점하느냐의 싸움이다. 기억이 없는 에이전트는 매번 자기소개를 다시 해야 하는 낯선 타인일 뿐이며, 이런 도구는 결코 우리의 삶을 혁신하는 비서가 될 수 없다.

둘째와 셋째는 시각과 상식이다. Meta의 SAM 3로 얻은 입체적인 눈과, SIMA 2로 얻은 물리적인 뇌는 에이전트를 모니터 밖의 세상으로 이끌 것이다. 텍스트로 배운 깨진 유리잔이 아니라, 시뮬레이션 속에서 수천 번 깨뜨려보며 체득한 감각만이 로봇이 인간과 안전하게 공존할 수 있는 자격을 부여한다. 이것이 갖춰질 때, 에이전트는 비로소 가상의 비서를 넘어 물리적 현실을 제어하는 로봇으로 확장될 수 있다.

이 세 가지 요소는 독립적이지 않다. **눈으로 본 것을 기억에 저장하고, 그 기억을 바탕으로 뇌가 판단한다. 그중에서도 첫 단추는 단연 기억이다**. 기억이 없는 눈은 매 순간 새로운 세상을 보는 것이고, 기억이 없는 뇌는 매 순간 처음부터 다시 생각해야 하기 때문이다. B2B에서는 업무의 흐름을, B2C에서는 소비의 맥락을 기억하는 모델이 다가올 에이전트 시장의 패권을 쥐게 될 것이다. 다시 한번, 우리는 지금 단순한 연산 능력을 가진 지능의 시대에서 맥락과 경험을 기억하는 지성과 지혜의 시대로 넘어가고 있음을 명심하자.

III. AI Peer Table

표 8. 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	277.55	6,024	3.3	5.7	21.2	42.4	18.6	11.3	01-30	엔드유저 데이터 보유 및 서비스 배포능력
테슬라	TSLA US	426.58	2,084	5.6	-1.6	21.3	25.7	26.1	5.6	01-29	FSD v13 공공 출시 및 1Q25 중국 출시 가능성
Pony AI	PONY US	13	9	12.6	-27.1	-9.2	-22.7	3.5	-6.3	03-25	중국의 자율주행 선도기업 중 하나
삼성전자	005930 KS	103,400	613	2.1	1.8	47.2	94.0	88.1	97.8	01-07	온디바이스 AI, HBM, 파운드리 영역에서의 가능성
듀오링고	DUOL US	188.43	13	8.0	-39.6	-40.4	-63.8	-46.7	-41.9	02-27	AI 캐릭터와 상호작용하는 학습 플랫폼
크래프톤	259960 KS	256,500	12	1.0	-10.2	-24.3	-32.7	-17.4	-17.9	02-11	PUBG, inZOI 등 자사 게임에 AI CPC 기능 탑재
쇼피파이	SHOP US	159.34	305	9.1	-7.9	13.1	57.0	42.2	49.9	02-11	이커머스 콘텐츠 생성의 자동화(Shopify Magic AI)
코인베이스	COIN US	264.97	105	3.0	-25.2	-14.1	0.7	-9.7	6.7	02-13	AI 에이전트간 자율 암호화폐 거래 인프라 구축
비즈니스 효율화											
크라우드스트라이크	CRWD US	501.54	185	-3.7	-4.9	20.1	10.1	37.7	46.6	12-02	기업 고객 대상 엔드포인트 보안(Charlotte AI)
IBM	IBM US	303.21	416	5.1	-0.8	25.6	18.7	36.0	41.5	01-28	기업 고객 대상 AI 모델 개발, 배포 플랫폼(watsonx)
액센츄어	ACN US	247.85	227	2.8	0.1	-2.3	-19.0	-30.3	-28.0	12-18	각 산업별 맞춤형 AI 에이전트를 만들기 위한 첨병
SAP	SAP GY	206.70	432	0.0	-12.2	-12.6	-22.2	-5.3	-11.7	01-29	기업용 SW 솔루션에 AI를 내장한 에이전트(Joule)
인포시스	INFO IN	1,564.00	107	1.8	4.0	3.8	2.6	-16.3	-14.3	01-14	각 산업별 맞춤형 AI 에이전트를 만들기 위한 첨병
일루미나	ILMN US	129.81	29	8.1	29.7	31.5	61.5	-9.2	-2.9	02-06	AI 기반 첨단 유전체 분석
HD현대	267250 KS	201,500	16	-5.6	-5.0	50.0	134.9	163.0	164.0	02-06	AI, 로보틱스, 디지털 트랜스포메이션 관련 노력
AI Ops											
클라우드플레어	NET US	196.99	102	0.4	-9.6	-0.8	24.6	92.7	82.9	02-06	실시간 AI 추론에 CDN 수요증가 가능성(Workers AI)
팔란티어	PLTR US	165.77	580	0.2	-10.2	3.0	34.4	152.2	119.2	02-03	비즈니스 현장의 의사결정을 돕는 AI 플랫폼(AIP)
파운데이션 모델											
알파벳	GOOGL US	319.95	5,674	9.3	23.1	54.6	90.3	90.0	69.6	02-04	OpenAI와의 AGI 경쟁이 가능한 업체(Gemini)
메타 플랫폼스	META US	633.61	2,346	7.3	-14.2	-15.9	1.2	10.8	8.5	01-29	오픈소스 AI 개발의 선구자(Llama 등)
알리바바	9988 HK	151.70	547	-2.0	-12.6	24.9	30.7	81.4	87.2	02-20	중국의 CSP이자 중국의 대표적인 언어모델(Qwen)
네이버	035420 KS	255,000	40	-1.7	1.4	17.0	38.7	26.3	28.9	02-06	한국형 LLM 개발 및 AI 서비스(HyperCLOVA)

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

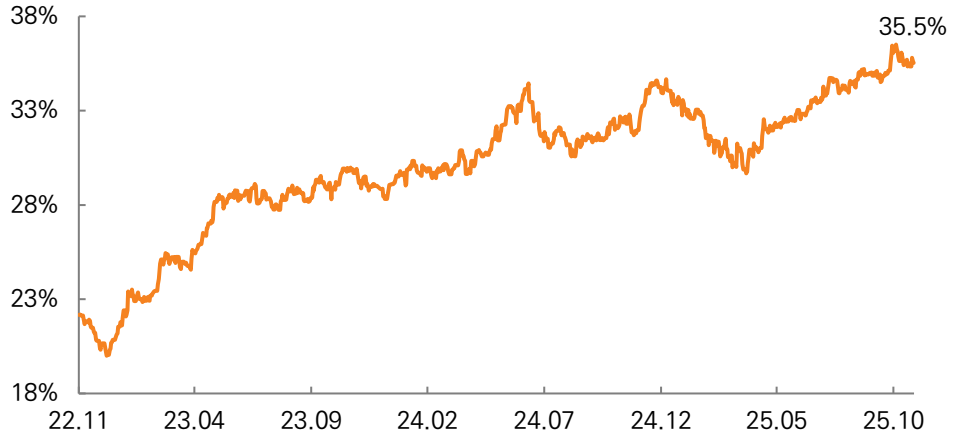
표 9. Peer Table (2)

종목명	티커	주가 (현지통화)	시가총액 (조원)	기간별 수익률(%) 1W	1M	3M	6M	1Y	YTD	실적발표 예정일	AI 밸류체인 관련 한줄평
클라우드											
마이크로소프트	MSFT US	485.50	5,301	-0.1	-7.1	-3.1	8.2	14.3	16.0	01-29	OpenAI 모델 라이선스 독점권 보유한 세계 2위 CSP
아마존 닷컴	AMZN US	229.16	3,599	2.9	2.2	0.2	14.0	10.2	4.5	02-06	최고의 CSP로서 Anthropic에 수 조원 투자
오라클	ORCL US	204.96	858	-9.1	-27.7	-12.3	31.9	8.8	24.2	12-09	AI 데이터센터 capex 경쟁에 진입한 전통적 강자
소프트뱅크	9984 JP	16,615.00	223	-13.4	-34.8	7.9	113.2	85.7	81.6	02-12	'비전펀드'는 AI에 집중. 일본 최고의 AI 슈퍼컴퓨터.
하드웨어 인프라											
엔비디아	NVDA US	180.26	6,434	-3.4	-3.2	-0.8	37.3	31.7	34.3	02-26	AI 모델 훈련 및 추론에 필수인 GPU계 현존 최강자
브로드컴	AVGO US	397.57	2,758	12.2	12.3	33.6	74.5	143.7	72.7	12-11	이더넷 기반 네트워킹 반도체의 최강자
AMD	AMD US	214.24	512	-4.2	-15.3	28.6	94.2	55.6	77.4	02-04	서버용 CPU의 최강자. AI 가속기 분야 패스트팔로어
인텔	INTC US	36.81	258	4.8	-3.8	51.2	83.6	53.1	83.6	01-30	파운드리로서 소버린 AI 미 지정학적 가치 부상
마이크론	MU US	230.26	381	1.9	5.1	97.8	147.0	127.1	174.4	12-18	SK하이닉스 추격 중인 "미국"의 메모리 반도체
SK하이닉스	000660 KS	539,000	392	-5.0	-2.5	107.7	167.2	223.6	213.2	01-22	HBM 부문 전세계 최강자
시놉시스	SNPS US	409.68	112	6.1	-11.7	-31.3	-17.9	-26.0	-15.6	12-10	반도체 EDA 부문 리딩 업체, AI 에이전트 적극 활용
케이던스 디자인 시스템	CDNS US	306.35	122	-0.1	-11.2	-11.0	-2.9	-1.2	2.0	02-18	반도체 EDA 부문 리딩 업체
암페놀	APH US	138.72	249	1.5	3.7	26.4	62.7	90.9	100.9	01-28	엔비디아향 AI용 서버 내 고속 I/O 및 커넥터 제조
버티브	VRT US	172.02	97	0.8	-7.5	34.6	65.3	30.8	51.6	02-12	액체 냉각 방식에 있어 글로벌 선두업체
Arm	ARM US	132.61	207	-3.2	-22.3	-5.5	4.3	-3.0	7.5	02-05	AI 가속기 설계 위한 다수의 IP를 소유한 팹리스
마벨 테크놀로지스	MRVL US	87.72	111	7.9	4.3	18.2	44.8	-5.5	-20.3	12-02	데이터 인프라용 네트워킹 및 스토리지 ASIC 설계
코어위브	CRWV US	74.29	55	-0.8	-44.0	-18.7	-27.7	-	-	02-13	AI용 클라우드 인프라 제공업체
중과서광	603019 CH	100.71	31	0.8	-12.4	13.5	62.7	28.9	39.9	03-04	중국과학원과 긴밀 관계를 유지 중인 HPC 제조업체
SMIC	981 HK	69.75	126	-5.1	-15.8	24.0	66.7	171.4	119.3	02-11	7나노 공정을 달성한 중국 유일의 미세공정 파운드리
TSMC	2330 TT	1,440.00	1,750	-1.0	-2.7	21.5	50.5	46.4	35.7	01-16	명실상부 반도체 파운드리 최강자
폭스콘	2317 TT	229.50	150	-3.0	-6.9	10.3	56.4	21.0	29.2	03-16	GPU 기반 보드 및 서버 시스템 조립 담당, 대만 업체
위스트론	3231 TT	144.50	22	1.0	-1.4	25.7	30.0	32.9	43.8	01-16	GPU 기반 보드 및 서버 시스템 조립 담당, 대만 업체
Alchip	3661 TT	3,295.00	13	7.2	7.5	-18.5	16.3	54.0	1.4	03-02	빅테크들의 AI 가속기 설계를 돕는 대만의 팹리스
에너지 인프라											
GE 베르노바	GEV US	589.72	235	-0.9	0.9	-5.7	27.1	74.4	79.6	01-22	AI 데이터센터용 가스터빈 인프라 솔루션
슈나이더 일렉트릭	SU FP	227.30	223	-0.5	-12.2	7.5	2.0	-2.7	-4.0	02-19	AI 데이터센터 전력관리 및 냉각시스템 통합 솔루션
비스트라	VST US	176.80	88	-1.3	-12.2	-9.3	12.0	9.3	28.8	02-27	미국 최대 독립 전력 생산 사업자
블룸 에너지	BE US	101.14	35	-7.2	-8.4	102.5	419.2	286.5	355.4	02-27	AI 데이터센터용 연료전지 솔루션
넥스트에라 에너지	NEE US	85.54	262	2.2	2.0	16.0	29.1	13.8	23.0	01-23	AI 데이터센터용 엔드투엔드 전력 파트너

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

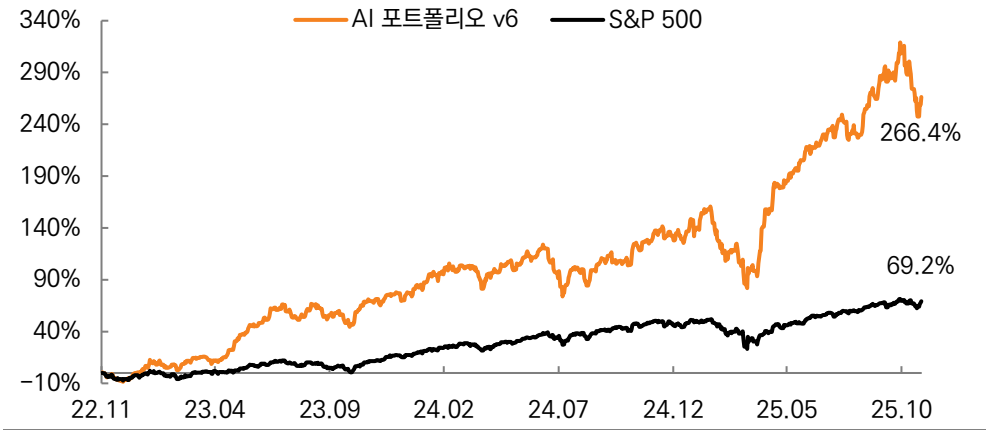
IV. Appendix: Charts

그림 42. ChatGPT 출시일 이후, S&P 500에서 Magnificent Seven이 차지하는 비중(시가총액 기준)



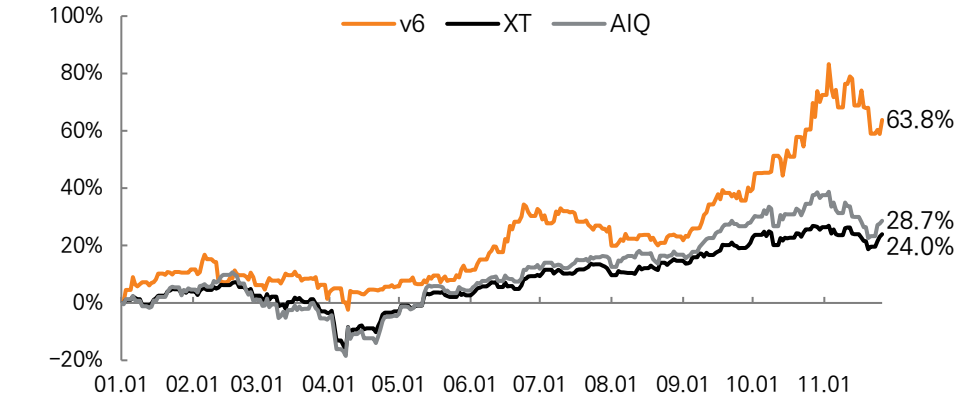
자료: Bloomberg, 미래에셋증권 리서치센터

그림 43. ChatGPT 출시일 이후, AI 밸류체인에 속한 유니버스 종목들의 평균 수익률(동일가중 방식)



자료: Bloomberg, 미래에셋증권 리서치센터

그림 44. 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 10. 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(New)	Gemini 3 Pro Preview	76.4	구글	독점 폐쇄소스	2025년 1월
2(▼1)	Gemini 2.5 pro(06-05)	62.4	구글	독점 폐쇄소스	2025년 1월
3(New)	Claude Opus 4.5	62.0	Anthropic	독점 폐쇄소스	2025년 8월
4(▼2)	GPT-5 Pro	61.6	OpenAI	독점 폐쇄소스	2024년 9월
5(▼2)	Grok 4	60.5	xAI	독점 폐쇄소스	2024년 12월
6(▼2)	Claude 4.1 Opus	60.0	Anthropic	독점 폐쇄소스	2025년 3월
7(▼2)	Claude 4 Opus	58.8	Anthropic	독점 폐쇄소스	2025년 3월
8(▼2)	GPT-5 high	56.7	OpenAI	독점 폐쇄소스	2024년 9월
9(New)	Grok 4.1 Fast	56.0	xAI	독점 폐쇄소스	-

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

표 11. 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(New)	gemini-3-pro	1507	구글	독점 폐쇄소스	2025년 1월
2(New)	claude-sonnet-4-5-20251101-thinking-32k	1496	Anthropic	독점 폐쇄소스	2025년 7월
3(New)	claude-opus-4-5-20251101	1494	Anthropic	독점 폐쇄소스	2025년 8월
4(New)	grok-4.1-thinking	1487	xAI	독점 폐쇄소스	-
5(▼4)	claude-sonnet-4-5-20250929-thinking-32k	1481	Anthropic	독점 폐쇄소스	2025년 3월
6(▼4)	claude-opus-4-1-20250805-thinking-16k	1480	Anthropic	독점 폐쇄소스	2025년 3월
7(New)	gpt-5.1-high	1475	OpenAI	독점 폐쇄소스	2024년 9월
8(New)	grok-4.1	1473	xAI	독점 폐쇄소스	-
9(▼6)	claude-sonnet-4-5-20250929	1472	Anthropic	독점 폐쇄소스	2025년 7월
10(▼6)	claude-opus-4-1-20250805	1470	Anthropic	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 12. LLM 벤치마크 LiveBench의 성능 상위 모델 10개(동일 모델 생략)

순위 (전 주 변동)	모델명	전체 평균 점수	기관명	라이선스	지식 컷오프
1(New)	Claude 4.5 Opus Thinking High Effort	75.58	Anthropic	독점 폐쇄소스	2025년 7월
2(New)	Claude 4.5 Opus Thinking Medium Effort	74.87	Anthropic	독점 폐쇄소스	2025년 7월
3(New)	Gemini 3 Pro Preview High	74.14	구글	독점 폐쇄소스	2025년 1월
4(▼3)	GPT-5 high	73.51	OpenAI	독점 폐쇄소스	2024년 9월
5(▼2)	GPT-5 Pro	73.48	OpenAI	독점 폐쇄소스	2024년 9월
6(▼1)	GPT-5 Codex	73.36	OpenAI	독점 폐쇄소스	2024년 9월
7(New)	GPT-5.1 High	72.52	OpenAI	독점 폐쇄소스	2024년 9월
8(▼6)	GPT-5 medium	72.26	OpenAI	독점 폐쇄소스	2024년 9월
9(▼5)	Claude Sonnet 4.5 Thinking	71.83	Anthropic	독점 폐쇄소스	2025년 7월
10(New)	GPT-5.1 Codex	70.84	OpenAI	독점 폐쇄소스	2024년 9월

자료: LiveBench, Huggingface, 미래에셋증권 리서치센터

표 13. 코딩 관련 벤치마크 SWE-Bench의 성능 상위 모델 10개

순위 (전주 변동)	모델명	점수(%)	기관명	등록일
1(New)	Claude 4.5 Opus medium(20251101)	74.40	Anthropic	2025-11-24
2(New)	Gemini 3 Pro Preview(20251118)	74.20	구글	2025-11-18
3(▼2)	Claude 4.5 Sonnet (20250929)	70.60	Anthropic	2025-09-29
4(▼2)	Claude 4 Opus (20250514)	67.60	Anthropic	2025-08-02
5(▼2)	GPT-5 (2025-08-07) (medium reasoning)	65.00	OpenAI	2025-08-07
6(▼2)	Claude 4 Sonnet (20250514)	64.93	Anthropic	2025-05-21
7(▼2)	GPT-5 mini (2025-08-07) (medium reasoning)	59.80	OpenAI	2025-08-07
8(▼2)	o3 (2025-04-16)	58.40	OpenAI	2025-05-21
9(▼2)	Qwen3-Coder 480B/A35B Instruct	55.40	Alibaba	2025-08-02
10(▼2)	Gemini 2.5 Pro (2025-05-06)	53.60	구글	2025-05-21

자료: SWE-Bench, 미래에셋증권 리서치센터

표 14. 컨텍스트 메모리 관련 벤치마크 LongBench의 성능 상위 모델 10개

순위 (전주 변동)	모델명	기관명	모델 크기	컨텍스트 길이	출시일	종합 성능	긴 글 성능(%)
-	Human	-	-	-	-	53.7	53.7
1(-)	Gemini-2.5-Pro	구글	-	1M	2025-03-25	63.3	71
2(-)	Gemini-2.5-Flash	구글	-	1M	2025-04-17	62.1	55.7
3(-)	Qwen3-235B-A22B-Thinking-2507	알리바바	235B	256k	2025-07-25	60.6	58.1
4(-)	DeepSeek-R1	DeepSeek	671B	128k	2025-01-20	58.3	59.3
5(-)	Qwen3-235B-A22B-Instruct-2507	알리바바	235B	256k	2025-07-22	58.3	55.6
6(-)	o1-preview	OpenAI	-	128k	2024-09-12	57.7	58.1
7(-)	DeepSeek-R1-0528	DeepSeek	671B	128k	2025-05-28	56.7	51.4
8(-)	MiniMax-Text-01	MiniMax	456B	4M	2025-01-15	56.5	47.2
9(-)	Gemini-2.0-Flash-Thinking	구글	-	1M	2025-01-21	56.0	49.1

자료: LongBench, 미래에셋증권 리서치센터
주: LongBench는 멀티태스크·이중언어 롱컨텍스트의 클래식. 실제 과업 분포가 좋아 모델 비교에 널리 쓰임. 긴 글 성능은 128K 기준

표 15. 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	특정 도메인
2(-)	Claude-3.7 Sonnet	56	42.11	2025년 2월	범용 에이전트
3(-)	o3 Medium	54	14.56	2025년 4월	특정 도메인
4(-)	Claude Opus 4.1	54	180.49	2025년 8월	범용 에이전트
5(-)	Claude-3.7 Sonnet High	52	31.94	2025년 2월	특정 도메인
6(-)	Claude Opus 4.1 high	52	149.98	2025년 8월	특정 도메인
7(-)	Claude Opus 4.1	50	69.78	2025년 8월	특정 도메인
8(-)	GPT-5 medium	48	23.83	2025년 8월	특정 도메인
9(-)	DeepSeek V3	44	5.43	2025년 3월	특정 도메인
10(-)	Claude-3.7 Sonnet	44	15.45	2025년 2월	특정 도메인

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 알파벳 40주 보유하고 있습니다.
- 본 자료를 작성한 애널리스트 김은지은(는) 자료작성일 현재 엔비디아 60주 보유하고 있습니다.
- 본 자료를 작성한 애널리스트 한종목은(는) 자료작성일 현재 루멘텀 홀딩스 8주 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.