

AI NEWS

251013-251019

가짜연구소

| SOTA AI Review Team

| 작성자 : 송건학

목 차

10위 Andrej Karpathy 100달러로 구축 가능한 모델
'NanoChat' 공개

9위 코딩 AI가 글쓰기보다 더 빨리 향상된 이유는
'강화 격차'

8위 Google, 동영상 모델 Veo 3.1 출시

7위 MS "Github, 데이터 저장소에서
Vibe Coding Platform으로 대대적 개편"

6위 xAI, '세계 모델' 경쟁에 합류.
"내년 말 AI 생성 게임 공개"

5위 Anthropic, Claude Agent Skills 공개

4위 Anthropic, 작고 저렴한 Claude Haiku 4.5 출시.
속도 2배·비용 1/3

3위 OpenAI, 1조달러(1400조원) 마련 위해 5년 계획
수립. 수익 다각화 전략

2위 OpenAI, Broadcom과 10GW 칩 계약

1위 MS, Windows 11 PC 전체에
음성비서·Agent 기능 도입

App. MS, MAI-Image-1 자체 첫 Text-to-Image Model 공개



[1] A roadrunner sprinting across sand [2] MAI-Image-1 written in the sand at sunset over the beach [3] A man crossing a city street

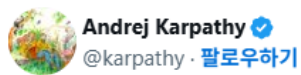


Text-to-Image			5 days ago
Rank (UB) ↑	Model ↓	Score ↓	
1	hunyuan-image-3.0	1161	🔍
1	gemini-2.5-flash-image-previ...	1154	
3	imagen-4.0-ultra-generate-pr...	1145	
3	seedream-4-2k	1144	
4	seedream-4-high-res-fal	1134	
5	imagen-4.0-generate-preview-...	1131	
7	gpt-image-1	1123	
7	seedream-4-fal	1118	
9	mai-image-1	1096	🔍
9	seedream-3	1082	

- MS는 현지시간 13일 Text-to-Image Generation Model 'MAI-Image-1' 공개함.
- 사진 수준의 사실적 이미지 생성이 강점이며, 번개와 자연 풍경 등 복잡한 장면을 높은 정밀도로 표현할 수 있다고 설명.
- 다만, 세부 성능이나 기술적인 부분 등은 공개하지 않음.
- 현재 LM Areana Text-to-Image Section 9위

<https://microsoft.ai/news/introducing-mai-image-1-debuting-in-the-top-10-on-lmarena/>

10. Andrej Karpathy 100달러로 구축 가능한 모델 'NanoChat' 공개



@karpathy · 팔로우하기

Excited to release new repo: nanochat!
(it's among the most unhinged I've written).

Unlike my earlier similar repo nanoGPT which only covered pretraining, nanochat is a minimal, from scratch, full-stack training/inference pipeline of a simple ChatGPT clone in a single, [더 보기](#)



오전 12:16 · 2025년 10월 14일

2.3만 답글하기 링크 복사하기

634개의 답글 읽기


- Andrej Karpathy가 개발한 Minimal LLM Project로 NanoChat 공개
- 목표 : 100달러로 ChatGPT를 만들어보자
(The best ChatGPT that \$100 can buy)
대규모 인프라 없이도 작동 가능한 저비용 LLM 구현을 실현하는 것
- 전형적인 ChatGPT 시스템의 모든 구성요소 (**tokenization, pretraining, finetuning, evaluation, inference, web serving**까지의 전체 파이프라인을 포함)를 하나의 코드베이스 안에 통합
- 권장 설정 : H100 8장을 4시간 동안 실행 (시간당 약 \$24, 총 \$100)
- 112억개의 토큰을 처리하는 5억6000만 매개변수의 모델 생성
- 학습 시간 12시간 확장시 1000달러(약 41.6시간 학습)까지 확장 가능
- 훨씬 더 일관성 있는 결과, 간단한 수학이나 코드 문제 가능
- Karpathy의 LLM 입문 강의 "LLM101n"의 핵심 과제가 될 예정

9. 코딩 AI가 글쓰기보다 더 빨리 향상된 이유는 '강화 격차'



- 코딩 분야는 몇 개월 전과는 비교가 안 될 정도로 발전했으며, 이는 Agent나 문서 작성 등의 더딘 발전 속도를 크게 앞서는 것
- 최근 몇 개월 간 **코딩 AI** 성능이 글쓰기와 같은 다른 기능 대비 빠르게 향상된 이유는 **강화 학습(RL)에 최적화된 구조** 때문
- RL은 인간이 목표를 달성하기 위해 사용하는 시행착오 학습 과정을 모방한 ML 기법으로, 데이터를 처리할 때 Reward Mechanism을 사용해 피드백을 얻고 이를 통해 최상의 처리 경로를 스스로 발견하는 구조.
- 코딩 분야는 명확한 답이 있어, 결과 측정 및 Reward Mechanism에 있어 RL 최적화 가능하지만, 글쓰기 등은 모델의 답이 정답인지를 가릴 수 있는 객관적인 지표가 없음. 이에 따라서 코딩이나 수학에 비해 발전이 더디다는 것
- 여전히 일부 결과물이 불완전하고 시스템과의 통합, 보안 등에서 문제가 발생하는 경우가 많지만, 빠르게 보완될 가능성이 높음.
- 강화 격차(Reinforcement Gap)가 심한, 즉 AI 성능이 빠르게 발전하는 분야는 그만큼 인간 대체가 빨라질 것으로 분석

8. Google, 동영상 모델 Veo 3.1 출시 Sora 2 평가 엇갈려



Veo 3.1

Veo 모델에 관해 자세히 알아보세요

대한민국

Veo 3.1 Fast

고화질은 유지하며 속도를 최적화하는 동영상 생성 모델로 사운드가 포함된 동영상을 만들어 보세요.

Google AI Pro 요금제 사용

지금 구독

- ✓ 8초 분량의 동영상을 만들어 보세요
- ✓ 고화질, 빠른 속도를 위해 최적화됨

신규 네이티브 오디오 생성

Veo 3.1

최첨단 동영상 생성 모델을 사용하여 사운드가 포함된 8초 분량의 고화질 동영상을 만들어 보세요.

Google AI Ultra 요금제 사용

지금 구독

- ✓ 8초 분량의 동영상을 만들어 보세요
- ✓ 최고 수준의 동영상 품질

신규 네이티브 오디오 생성

- Google에서 Veo 3.1 출시
(25.05 Google I/O에서 Veo 3 공개. 25.09.30 Sora 2 공개)
- 영상의 물리적 표현과 질감을 크게 개선, AI로 생성된 영상의 조명·그림자·물리 효과를 세밀하게 조정할 수 있으며, 영상 속 오브젝트를 자연스럽게 추가하거나 제거 가능, 배경 재구성 가능
- 가장 큰 변화는 AI 오디오 생성 기능의 전면 통합.
 - 사용자는 영상과 음향을 동시에 생성하거나 자연스럽게 확장 가능
- 기능 통합으로 사용자는 별도의 후반 작업 없이도 톤과 감정, 스토리텔링이 유기적으로 어우러진 영상을 손쉽게 제작할 수 있게 됐다는 설명
- Standard 모델은 초당 0.40달러, Fast 모델은 초당 0.15달러로 책정
- 영상 해상도 : 720p~1080p, 24fps /
길이 : 기본 4~8초 (Extend 기능을 통해 최대 2분30초 이상)
- 업계 반응은 긍정적인 편이지만, 일부 전문가는 Sora 2에 못 미친다고 평함. 일부 Veo 3.1은 훌륭하지만, 품질 면에서는 Sora 2보다 다소 인공적이며 가격도 비싸다는 의견도 존재

7. MS “Github, 데이터 저장소에서 Vibe Coding Platform으로 대대적 개편”



Introducing the GitHub Copilot coding agent (25.06)

<https://www.youtube.com/watch?v=EPyyyB23NUU&t=1s>

- MS가 AI 코딩 도구 시장 경쟁 심화에 따라 Github를 대대적 개편 논의함.
- 내부 회의에서 유출된 발언을 인용, 사티아 나델라 CEO가 AI가 문서, 앱, 웹사이트의 경계를 허물고 있기에 Github를 AI 기반 소프트웨어 개발의 중심 플랫폼으로 재편하겠다는 구상을 보도
- 특히, AI 시대 앱과 문서, 웹사이트의 차이가 더 이상 의미가 없다고 강조
- MS는 AI 코딩 어시스턴트 GitHub Copilot을 개발자가 모든 환경에서 접근 가능하도록 확장할 계획.
- 웹 브라우저나 VS 코드 같은 개발 환경은 물론, CLI나 MS의 다른 제품 안에서도 Github의 AI 도구를 사용할 수 있도록 할 예정.
- Github를 여러 AI 에이전트를 통합 관리하는 대시보드 형태로 발전 구상
- 제이 파릭 MS 부사장은 “Github는 더 이상 단순히 코드를 저장하는 공간이 아니다. AI 중심의 소프트웨어 개발 생태계의 중력 중심으로 만들 것”
- Github Actions 기반 보안·분석 도구, 국가별 데이터 규제 준수 인프라 등을 강화해 서비스 품질을 높ی겠다고 덧붙임.

6. xAI, '세계 모델' 경쟁에 합류..."내년 말 AI 생성 게임 공개"

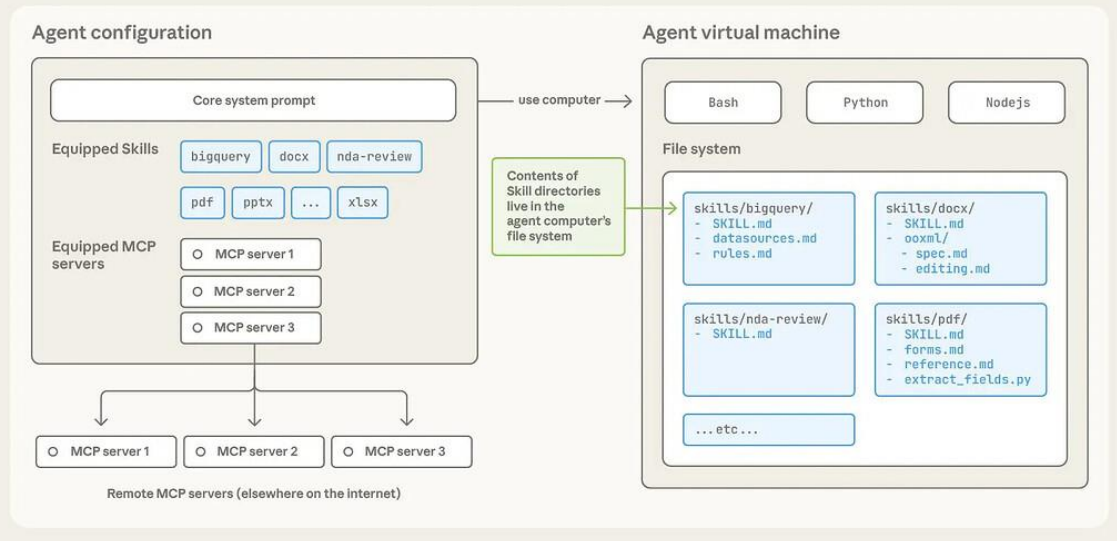


- Elon Musk는 xAI가 현실 세계를 이해하고 재현할 수 있는 차세대 AI 시스템인 World Model 개발에 속도를 내고 있다. 이를 통해 AI 생성 게임을 만들겠다는 목표 언급
- NVIDIA는 Omniverse 플랫폼을 통해 가상 환경에서 물리 기반 시뮬레이션을 구현해왔으며, 이 기술을 월드 모델의 핵심으로 꼽음
- **동영상 생성 모델**은 학습 데이터에서 도출한 패턴을 기반으로 프레임을 예측하는 수준. 반면, **월드 모델**은 사물이 실제 환경에서 어떻게 움직이고 상호작용하는지 이해하고 예측하는 것이 가능.
이는 단순 생성이 아니라, 물리적 인과 추론이 가능한 AI라는 점에서 차별화.
- xAI는 Omni 팀이라는 신규 조직 신설, 이미지·비디오·오디오 등 멀티모달 AI 경험을 설계할 인재를 채용 중. (연봉은 18만~44만달러(약 2억6000만~6억원)) AI를 테스트하는 '비디오 게임 튜터'도 시급 45~100달러로 모집 중
최근 엔비디아 출신 핵심 연구진을 영입, 월드 모델 전담팀을 구성
- xAI는 최근 Grok의 이미지·영상 생성 모델 Imagine v0.9 버전으로 업데이트
- World Model과 관련하여 로봇 등을 통해 촬영된 실제 환경 영상 데이터로 학습, 물리 법칙과 인과관계를 이해하도록 설계된 모델을 개발 계획

5. Anthropic, Claude Agent Skills 공개

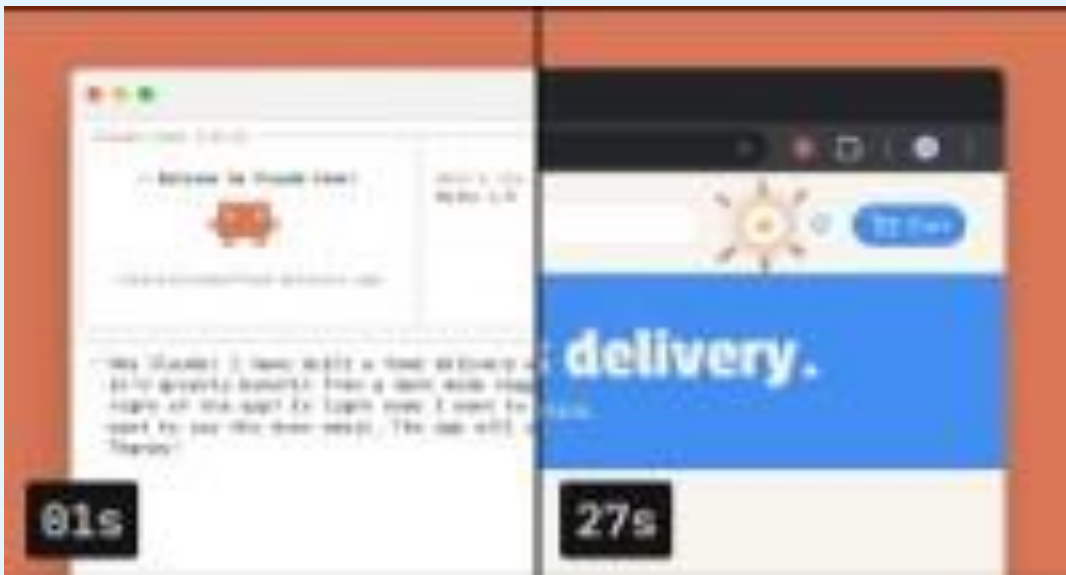


Agent + Skills + Virtual Machine



- Anthropic [Claude Skills](#) 를 공식 발표 (25.10.16)
- 모델이 특정 작업(예: Excel 작업, 조직의 가이드라인 준수) 수행 시, 필요한 지침, 스크립트, 리소스를 담은 **폴더 단위의 능력 확장 시스템**
- 작업과 관련이 있을 때만 해당 스킬에 접근하여 전문화된 작업 수행
- 구독 : 현재 Pro, Max, Team, Enterprise 사용자 모두 스킬 기능 사용 가능
- API : Messages API 요청과 새로운 /v1/skills 엔드포인트를 통해 커스텀 스킬의 버전 관리 및 운용 제어 가능
- 세션 시작 시 모든 사용 가능한 스킬 파일을 스캔하고 각 스킬의 **YAML의 짧은 설명만 읽음**
- 각 스킬이 차지하는 초기 토큰은 **수십 개에 불과**하여 극도로 효율적
- MCP의 주요 한계(토큰 사용량)를 벗어나 Skills를 통한 단순성과 공유의 용이성이 앞으로의 활용에 있어 두각을 나타낼 것으로 보임
- Anthropic 제공 자료 [Agent Skills 문서](#) [Claude Skills Cookbook](#)

4. Anthropic, 작고 저렴한 Claude Haiku 4.5 출시. 속도 2배·비용 1/3



	Claude Sonnet 4.5	Claude Haiku 4.5	Claude Sonnet 4	GPT-5	Gemini 2.5 Pro
Agentic coding <i>SWE-bench Verified</i>	77.2%	73.3%	72.7%	72.8% <small>GPT-5 (high)</small> 74.5% <small>GPT-5-Codex</small>	67.2%
Agentic terminal coding <i>Terminal-Bench</i>	50.0%	41.0%	36.4%	43.8%	25.3%
Agentic tool use <i>t2-bench</i>	Retail 86.2%	Retail 83.2%	Retail 83.8%	Retail 81.1%	—
	Airline 70.0%	Airline 63.6%	Airline 63.0%	Airline 62.6%	—
	Telecom 98.0%	Telecom 83.0%	Telecom 49.6%	Telecom 96.7%	—
Computer use <i>OSWorld</i>	61.4%	50.7%	42.2%	—	—
High school math competition <i>AIME 2025</i>	100% (python)	96.3% (python)	70.5%	99.6% (python)	88.0%
	87.0% (no tools)	80.7% (no tools)		94.6% (no tools)	
Graduate-level reasoning <i>GPQA Diamond</i>	83.4%	73.0%	76.1%	85.7%	86.4%
Multilingual Q&A <i>MMMLU</i>	89.1%	83.0%	86.5%	89.4%	—
Visual reasoning <i>MMMU (validation)</i>	77.8%	73.2%	74.4%	84.2%	82.0%

- Anthropic : 속도와 비용 효율을 크게 개선한 초소형 모델 Claude Haiku 4.5 공개
- 지난 5월 출시된 Claude Sonnet 4와 비슷한 성능을 제공하면서 속도는 2배 이상 빠르고 비용은 3분의 1 수준으로 낮춤
- 지연 시간이 중요한 실시간 서비스와 개발 환경에 최적화. 특히, 코드 생성 도구 Claude Code에서 활용하면 반응 속도가 크게 향상, 실시간 협업과 신속한 프로토타이핑이 쉬워진다는 설명
- SWE-Bench Verified에서 73.3%, Terminal-Bench에서 41% 기록, Sonnet 4.5보다는 다소 낮지만 Sonnet 4, GPT-5, Gemini 2.5 Pro와 비슷한 수준
- 현재 Claude 무료 사용자 전원에게 서비스
- API 가격도 기존보다 크게 낮춤.
입력 토큰 100만개당 1달러, 출력 토큰 100만개당 5달러.
(Sonnet 4의 약 3분의 1, Opus 4.1의 5분의 1 수준)
- 최근 AI 코딩 시장에서는 OpenAI의 GPT-5-Codex가 추격
- 올해 말이나 내년 초에 Claude Opus 4.5를 공개할 계획

3. OpenAI, 1조달러(1400조원) 마련 위해 5년 계획 수립. 수익 다각화 전략

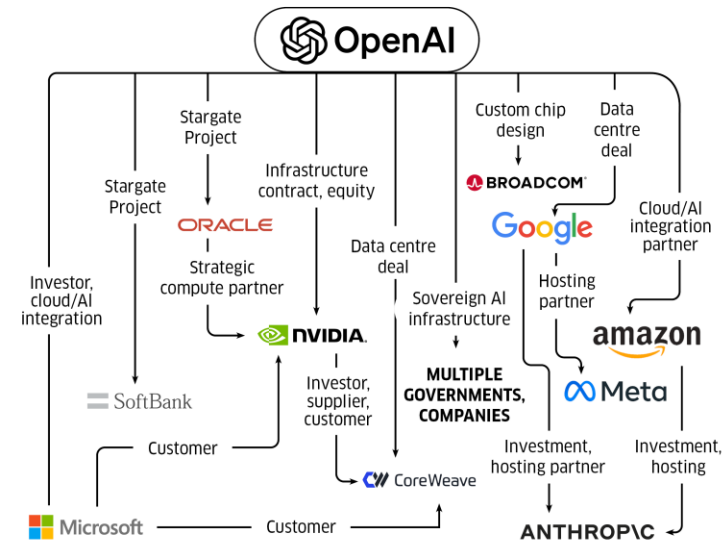
OpenAI Customers who have used 1T+ Tokens

Number	Name	Company	Role	Number	Name	Company	Role
1	Isaac Andersen	Duolingo	Senior SWE	16	Praty Sharma	HubSpot / Dashworks	AI / CoFounder
2	Alex Atallah	OpenRouter	CEO and CoFounder	17	Denis Shiryayev	JetBrains	Group Product Manager
3	Chris Colon	Indeed	Director, AI Platforms	18	Sam Spelsberg	Delphi	Co-founder & CTO
4	John Emmons	Salesforce	AI Leadership	19	Ashwin Sreenivas	Decagon	Co-founder
5	Harjot Gill	CodeRabbit	CEO and CoFounder	20	Shriram Sridharan	Rox	Co-founder
6	Cris Ippolite	iSolutionsAI	CEO and Director of AI	21	Nandan Thor	T-Mobile	VP of AI, Product & Engineering
7	Jiahui Jiang	Outtake	Engineering	22	Shashi Upadhyay	Zendesk	President, Product/Engineering/AI
8	Mahesh Kumar	Tiger Analytics	CEO and CoFounder	23	Aaron Weldy	Harvey	Software Engineer
9	Calvin Lee	Ramp	Founding Engineer	24	Luke Woloszyn	Read AI	Senior Data Scientist
10	Zachary Lipton	Abridge	Co-founder & CTO	25	Danny Wu	Canva	Head of AI Products
11	Joel Liu	Sider AI	Founder	26	Scott Wu	Cognition	Co-founder & CEO
12	Zach Lloyd	Warp.dev	Founder & CEO	27	Kai Xin Tai	Datadog	Product Manager
13	Dani Passos	Shopify	Developer/creator relations	28	Denis Yarats	Perplexity	Co-founder & CTO
14	Sarah Sachs	Notion	AI Lead / AI Engineering	29	Pablo Zamudio	Mercado Libre	AI & Data / ML Expert
15	Douglas Schonholtz	WHOOOP	Senior AI Engineer	30	Kay Zhu	Genspark AI	Co-founder & CTO

Startup
Scaled Company

- OpenAI가 1조 달러가 넘는 AI 인프라 구축을 위한 5년 계획에 돌입
- 1) 정부(B2G)와 기업(B2B) 대상 한 맞춤형 AI 솔루션 계약을 확대 예정. (현재 절반수준)
- 2) 소비자 대상 쇼핑 기능을 통한 수수료 및 Sora, AI Agent 등 프리미엄 서비스에서 매출을 창출할 계획 (8억명 사용자. 유료 가입자 5% -> 구독자 2배 증가 계획)
- 3) 새로운 AI 칩 개발에 따른 지적재산권(IP) 수익화, 온라인 광고 사업 진출, 현재 제작 중인 AI 하드웨어를 통한 매출 등도 거론
- 4) 최근 Oracle, NVIDIA, AMD, Broadcom 등과 계약 (26GW 규모의 컴퓨팅 용량을 확보, 1조달러). 파트너사가 초기 인프라 자본 지출 부담.
 - OpenAI는 장기적으로 운영 수익으로 이를 상환한다는 방침
- 현재 연간 반복 매출(ARR) 약 130억달러(약 18조원) 수준
 - 올해 상반기 매출 전년 대비 두배 이상, 영업손실 약 80억달러(약 11조원).
- 경영진 : 폭발적인 성장세가 이어진다면 추가 투자나 매출 증가도 충분히 가능 주장
- OpenAI는 앞으로 공급업체 경쟁과 기술 발전으로 컴퓨팅 비용이 급격히 하락 전망.
- 전체 컴퓨팅 인프라 개발 비용의 3분의 2는 반도체 조달에 투입.

LARGE TECH COMPANIES ARE TYING THEIR FORTUNES TO OPENAI



Source: Bloomberg News, Citi Research, FT Research

Addi. OpenAI : 12월부터 챗GPT에서 성인 콘텐츠 허용 예정. 일부 반발



Ok this tweet about upcoming changes to ChatGPT blew up on the erotica point much more than I thought it was going to! It was meant to be just one example of us allowing more user freedom for adults. Here is an effort to better communicate it:

As we have said earlier, we are [더 보기](#)



We made ChatGPT pretty restrictive to make sure we were being careful with mental health issues. We realize this made it less useful/enjoyable to many users who had no mental health problems, but given the seriousness of the issue we wanted to get this right.

Now that we have

오전 4:11 · 2025년 10월 16일

❤️ 6.2천 💬 답글하기 🔗 링크 복사하기

1.6천개의 답글 읽기

- Sam Altman은 14일 X를 통해 오는 12월부터 성인 인증을 완료한 사용자에게 성인용 에로틱 대화를 허용할 예정이라고 발표
- 기존 정신 건강 문제를 우려해 안전 장치를 유지해온 OpenAI가 처음으로 "성인 사용자는 성인처럼 대우한다"는 원칙을 내세우며 정책 전환
- 연령, 사용자가 직접 설정을 선택할 수 있는 필터, 법적, 윤리적 지침을 준수하는 한 모든 성인용 주제를 탐색할 수 있는 자유도 제공
- xAI는 3월 Grok에 업계 최초로 성인 모드를 공식 도입
- 성인 서비스는 사용자 참여를 높이고 수익에도 도움이 된다는 분석.
- 다만, 관련 이슈로 인한 비판 목소리에 Sam Altman은 15일 X를 통해 "OpenAI는 세계의 도덕 경찰로 선출된 것이 아니다"라고 밝힘. 6200개가 넘는 댓글이 달리는 등 뜨거운 찬반 토론 진행 중
- 미국의 비영리단체 전국성착취반대센터(NCOSE)를 비롯한 각계 단체들은 성명을 내고 오픈AI에 결정 철회를 촉구.

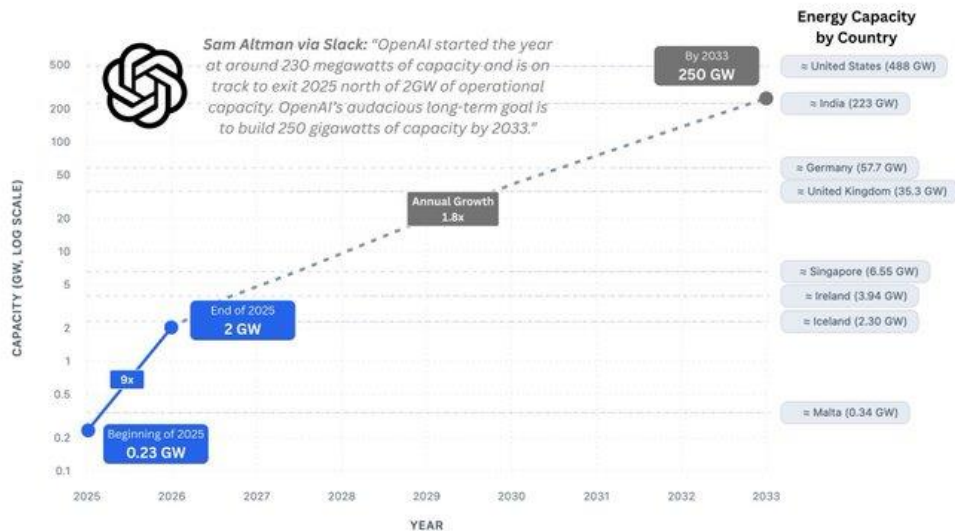
2. OpenAI, Broadcom과 10GW 칩 계약 (누적 26GW)



- OpenAI : 13일 Broadcom과 **10GW** 규모의 맞춤형 AI 칩 구매 계약 체결
- NVIDIA 및 AMD 파트너십 기반 **GPU & 자체 칩 Stargate** 프로젝트 투입
- **OpenAI : 칩 설계, Broadcom : 개발 & 배치** : 새 AI 칩은 Broadcom의 Ethernet 네트워크 기술을 기반으로, OpenAI의 AI 워크로드에 최적화된 메모리·연산·통신 기능을 통합한 구조로 개발.
- 10GW 규모의 AI 컴퓨팅 용량 구축시,
현재 칩 가격에 3500억달러(약 500조원),
인프라 구축까지 포함하면 총 5000억달러(약 700조원) 필요 전망
- 오픈AI가 확보할 총 AI 컴퓨팅 용량은 26GW.
 - 원자력 발전소 26기의 전력 규모, 1조 달러(약 1400조원) 이상이 소요 예상
- OpenAI가 2033년까지 목표로 잡은 250GW 구축에는 약 12조5000억달러(약 1경7843조원)가 필요할 것으로 추산

OpenAI planning to 125x energy capacity in 8 years

This would mean using more than India's energy capacity today



Source: Alex Heath, Sources.News
Peter Gostev (<https://x.com/petergostev>); (<https://www.linkedin.com/in/peter-gostev/>)

OpenAI x Broadcom – The OpenAI Podcast Ep. 8

<https://youtu.be/qqAbVTFnfk8?si=1TbEIUp3Lazu-rCu>

1. MS, Windows 11 PC 전체에 음성비서·Agent 기능 도입



Copilot on Windows 11 | Meet the Computer You Can Talk To

<https://youtu.be/KgQMchwq334?si=RHBCiyTZmcK5816u>

- MS가 16일 Windows 11 운영체제에 AI를 음성 기반 AI와 시각 인식 기능, 자율적 작업 수행 에이전트로 전면 개편한다고 발표
- 이를 통해 사용자와의 소통을 강화하고 Agent 기능을 결합한 차세대 AI PC 시대를 열겠다는 전략을 제안함.
- 사용자가 단순히 컴퓨터와 대화하듯 AI에게 업무를 맡길 수 있는 차세대 인간-컴퓨터 상호작용 환경 구축이 목표.
모든 윈도우 11 PC에서 작동
- Copilot Voice를 통해 음성을 마우스와 키보드에 이은 세번째 입력 수단으로 자리매김하겠다는 계획 ("**Hey, Copilot**" 호출)
- 가장 실험적인 Copilot Actions은 AI가 사용자를 대신해 PC 내 파일과 애플리케이션을 자동으로 조작.
현재 Windows Insiders 대상으로 테스트 중이며, 사진 라이브러리 정리와 문서 데이터 추출, 다단계 업무 처리 등 다양한 작업을 수행
- 게임 영역에서는 Gaming Copilot을 도입, ROG Xbox Ally 등 장치에서 음성으로 게임 질문에 답하고 전략을 안내

<https://www.aitimes.com/news/articleView.html?idxno=203231>