AI 뉴스와 용어
월·수·금 오전 7시

LAB METAL
어려운 AI, 쉽고 재미있게

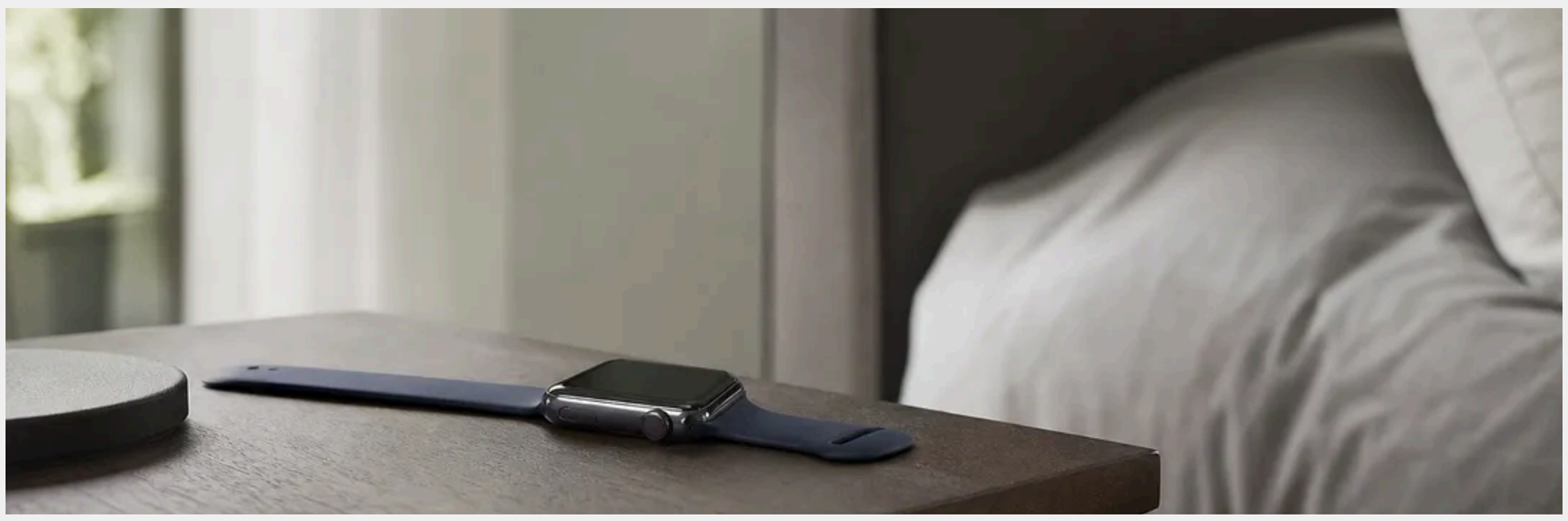
무료 · 언제든 수신거부
metallab.ai

제 10 호 | 2026년 8월 24일 월요일 | 오늘의 AI

비즈니스

AI 에이전트 24시간 굴리려 창업자들도 밤새운다

밤낮없이 일하는 AI 에이전트 속도를 못 따라가 창업자들이 잠을 줄이고 있어요



비즈니스 · AI 에이전트 24시간 굴리려 창업자들도 밤새운다

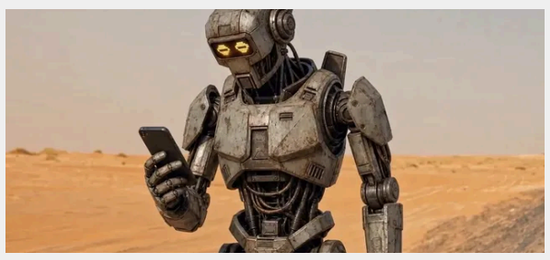
IT 문제 해결 소프트웨어를 만드는 Keel의 공동창업자 아디티야 샤르마(27)는 최근 생성형 AI 봇들과 밤을 새운 뒤 오전 6시에야 잠자리에 들었어요. 그가 부리는 에이전트들은 데이터를 뽑고 코드를 짜는 걸 넘어 신규 고객 기능을 만들고 사내 모델을 위한 리서치를 하고 마케팅 캠페인까지 지켜보고 있어요. 문제는 이 에이전트들이 작업 도중 막히면 사람의 승인이나 추가 정보가 필요하다는 점인데요, 그는 이 때문에 하루 24시간 대기 상태를 유지하려고 규칙적인 수면을 포기했다고 월스트리트저널과의 인터뷰에서 밝혔어요. "에이전트가 여덟 시간 막혀 있는 비용이 너무 크다"고요. 최근까지는 휴대폰이나 스마트워치로 에이전트를 원격 확인할 방법조차 없었다고 해요.

우리 용어사전에 따르면 AI 에이전트는 답변만 내놓는 챗봇과 달리 검색하고 파일을 만들고 예약까지 대신 처리하는 AI를 뜻해요. 여러 단계를 스스로 계획해서 실행하다 보니 사람이 잠든 시간에도 작업이 끝나거나 막힐 수 있고, 그때마다 다음 단계로 넘어가려면 사람의 확인이 필요해요. 게다가 에이전트를 움직이는 모델 성능이 몇 달 단위로 빠르게 바뀌면서 창업자들은 그때그때 업무 방식을 다시 짜야 하는 상황에 놓여 있어요.

30년 넘게 테크업계에 몸담고 회사를 네 번 성공적으로 매각한 페자리스는 올해 은퇴하겠다고 아내와 약속했었어요. 젊은 창업자 시절 무리한 근무로 두 차례 입원했고, 한 달 내내 하루 한 시간만 자다 사무실에서 쓰러진 적도 있다고 해요. 그런 그가 넉 달 전 다섯 번째 스타트업 Proxon을 차렸고, 개발자 여섯 명이 일하는 이 회사는 에이전트가 없을 때보다 서른 배 빠르게 돌아간다고 추정하고 있어요. 다만 에이전트가 일을 늘리면 사람의 일도 함께 늘어나요 — 온보딩 속도가 빨라진 만큼 고객 응대 요청도 늘어나는 식이죠.

데이터 수집 플랫폼 Riveter를 공동창업한 그릴스와 코디 와터스는 여전히 직원이 둘뿐이에요. 채용을 계속 고민하지만 그때마다 에이전트로 새 업무를 처리할 방법을 찾아내 채용을 미뤄왔다고 해요. 그릴스는 2024년 와이콤비네이터(Y Combinator) 액셀러레이터를 거치면서 같은 기수의 한 창업자가 고형식을 제대로 못 먹고 식사 대용품으로 버티다 입원하는 걸 지켜봤고요, 그 경험 때문에 지금은 쉴 때는 의식적으로 쉬려 한다고 밝혔어요. 와이콤비네이터 측은 창업자들에게 자기 몸을 돌보라고 오래전부터 권해왔다고 설명했어요.

보스턴에서 AIt를 운영하는 칼리아는 2023년 2월 스포티파이 AI그룹에서 일할 때 박사 학위를 받고 20년 경력을 쌓은 동료들이 "내 박사 학위가 쓸모없어졌다"고 말하는 걸 보고 AI가 세상을 뒤집을 거라 직감했다고 해요. 2024년 스포티파이를 나와 AIt를 세운 그는 최근 애플워치로 에이전트를 통제하기 시작했는데, 10분마다 워치가 진동하며 승인을 요청해와서 달리기 중에도 손목을 확인해야 하는 상황이 됐어요.



링크드인 AI슬롭 신고 버튼, 클릭 100만 건 돌파

페자리스는 에이전트와 일하는 걸 마약에 비유했어요. 실제로 마약을 해본 적은 없지만 아마 이런 느낌일 거라면서요. 그가 느끼는 감정의 바탕에는 일종의 포모(FOMO·기회를 놓칠지 모른다는 불안)가 있어요 — 일하지 않는 매 순간이 일종의 실패를 놓치는 셈이라는 계산이죠. 샤르마 역시

이 조사는 지난 7월 Epoch AI와 입소스(Ipsos)가 미국 직장인 1,106명을 대상으로 진행했는데, 20%가 동료나 외부 인력이 하던 업무를 AI에 맡기고 있다고 답했어요. 업무 유형별 AI 활용률은 기록 관리 25%에서 시스템·소프트웨어 설계 57%까지 벌어졌고요. 창업자들이 에이전트에 점점 더 많은 업무를 넘기면서도, 정작 사람이 감당해야 할 감시·승인·예외 처리 부담은 줄어들지 않고 있다는 게 이번 WSJ 취재가 보여주는 지점이에요.

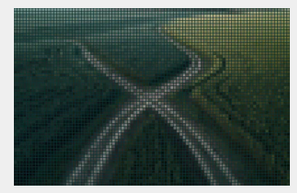
이번 취재가 보여주는 건 AI 에이전트가 일을 줄여주는 도구가 아니라 오히려 감독해야 할 새로운 노동력이 됐다는 사실이에요. 사람이 지시만 내리고 결과를 받아보는 그림은 아직 오지 않았고, 지금은 에이전트가 막힐 때마다 사람이 즉시 개입해야 하는 구조라 창업자의 노동시간이 줄기는커녕 오히려 24시간 대기 체제로 늘어난 거예요.

예전 스타트업의 '9-9-6' 문화(오전 9시부터 밤 9시까지 주 6일)와 비교하면 체감이 더 뚜렷해요. 그때는 사람이 할 수 있는 일의 양이 근무 시간의 한계였다면, 지금은 에이전트가 사람보다 먼저 일을 끝내버리기 때문에 창업자가 에이전트의 속도를 못 따라가는 게 새로운 병목이 됐어요. 페자리스가 추정하는 '30배 빠르다'는 체감은 과장이 아니라, 에이전트가 온보딩·코드 작성·모니터링을 동시에 처리하면서 사람에게 돌아오는 승인·검토 업무량 자체가 폭증했다는 뜻으로 읽는 게 맞아요.

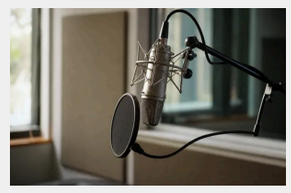
국내 스타트업이나 소규모 팀이 에이전트를 도입할 때도 참고할 지점이 여기 있어요. 에이전트를 늘릴수록 사람이 줄어드는 게 아니라, 에이전트가 막혔을 때 대신 판단해줄 사람의 대기 시간이 새로 생긴다는 걸 미리 계산에 넣어야 해요. 그릴스처럼 채용 대신 에이전트로 버티는 선택을 하기 전에, 승인·예외 처리에 하루 몇 시간이 들어가는지부터

오늘 지면

1 비즈니스 AI 에이전트 24시간 굴리려 창업자들도 밤새운다
2 카파시의 LLM 코딩 지적, CLAUDE.md 한 장으로 정리됐다
3 MCP 새 로드맵, 서버 상태 없이 수평 확장 지원
4 크리에이티브 힉스필드, 결혼식 선물 AI 러브스토리로 5억뷰 찍고 프롬프트 공개
5 RayNeo, 카메라 뺀 AI 안경 'iO 글래스' 공개
6 엔비디아, 하네스만 바꿔 AI 에이전트 점수 30%→100%로



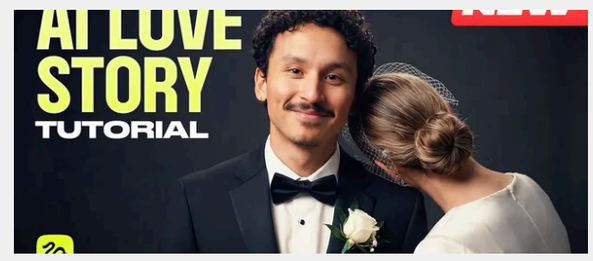
조지아텍, 올모3 완전공개로 사회적 추론의 뿌리 추적



ElevenLabs부터 세보니 '숫자+랩스' 이름이 줄줄이 나왔다

이 지면은 2026년 8월 24일 월요일 아침 구독자 메일함으로 나간 레터를 신문 판형으로 다시 조판한 것입니다. 기사 전문은 metallab.ai 에서 읽을 수 있습니다.

크리에이티브

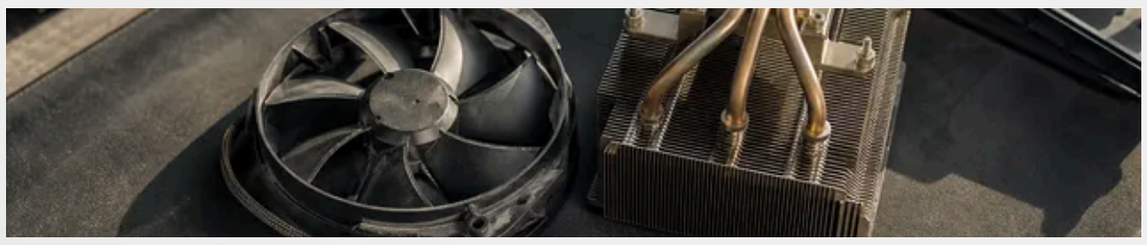


힉스필드, 결혼식 선물 AI 러브스토리로 5억 뷰 찍고 프롬프트 공개

힉스필드는 지인의 결혼식을 위해 만든 AI 러브스토리 단편이 소셜미디어 합산 약 5억 조회수를 기록했다고 유튜브 영상에서 밝혔다. 개인적인 축하 선물로 시작한 영상이 예상 밖의 반응을 얻자, 힉스필드는 이 작품의 전체 제작 과정과 프롬프트를 공개하는 튜토리얼 영상을 새로 올렸다. 영상 속 이야기는 여섯 개의 시대를 배경으로 두 명의 실존 인물이 등장하는 구조다. 항구, 경기장, 기차, 선박, 카니발, 엘리베이터까지 총 여섯 장소를 오가며 한 커

▶ 2면에 계속

카파시의 LLM 코딩 지적, CLAUDE.md 한 장으로 정리됐다

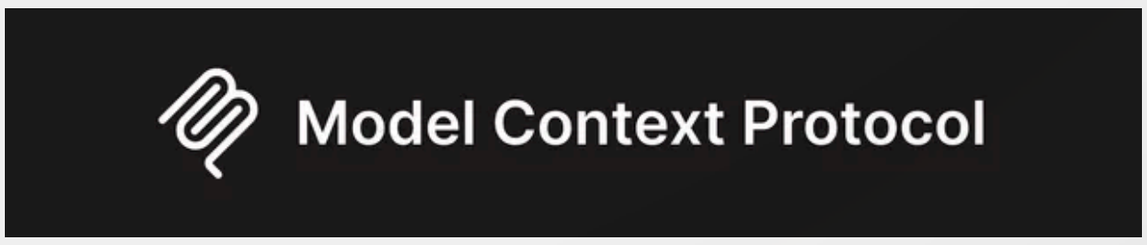


안드레이 카파시(Andrej Karpathy)가 엑스에 올린 글에서 LLM 코딩 에이전트가 반복하는 버릇을 지적했어요. multica-ai 저장소가 이 관찰을 바탕으로 클로드 코드(Claude Code)에 넣을 수 있는 지침을 만들고, "andrej-karpathy-skills"라는 이름으로 올려 뒀어요. 2026년 1월에 만들어진 이 저장소가 최근 깃허브 트렌딩에 올랐어요. 핵심 지침은 CLAUDE.md 한 장에 담겼지만, 저장소에는 설치와 다른 도구 연동을 위한 README·CURSOR.md·SKILL.md·플러그인·커서 규칙 파일도 함께 있어요. CLAUDE.md에는 카파시가 짚은 문제 세 가지에 대응하는 원칙 네 개가 담겼어요.

카파시는 LLM이 코드를 짤 때 세 가지를 반복한다고 지적했어요. "모델은 사용자 대신 잘못된 가정을 하고, 확인도 없이 그냥 진행해요." 혼란을 관리하지 않고, 명확히 해달라고

▶ 2면에 계속

MCP 새 로드맵, 서버 상태 없이 수평 확장 지원



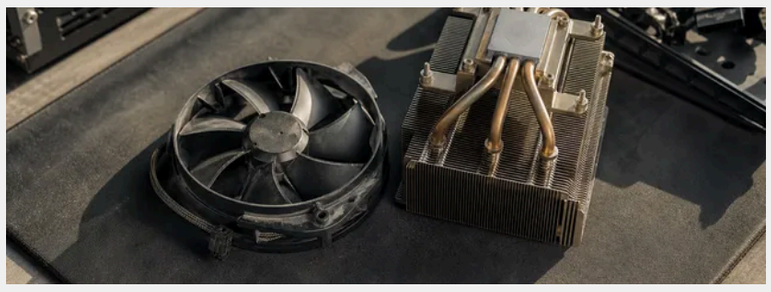
MCP(Model Context Protocol) 코어 메인테이너들이 업데이트된 로드맵을 공개했어요. AI를 외부 도구나 데이터에 연결하는 공용 규격인 MCP는 흔히 "AI계의 USB-C"로 불리는데, 이번 로드맵에서 가장 눈에 띄는 변화는 서버가 클라이언트와의 연결 상태를 계속 붙들고 있지 않아도 된다는 점이에요. SEP-2575와 SEP-2567을 통해 프로토콜 차원의 세션과 초기화 핸드셰이크가 사라졌고, 그 결과 서버가 상태를 갖지 않고도 수평 확장할 수 있게 됐다고 해요.

지난 3월 공개된 로드맵은 전송 진화와 확장성, 에이전트 통신, 거버넌스 성숙, 엔터프라이즈 준비라는 4개 우선순위를 제시했어요. 메인테이너들은 이 네 영역 모두에서 지난 5개월간 상당한 진전을 이뤘다고 밝혔는데요, 대부분의 변화

▶ 2면에 계속

◀ 1면에서 계속

카파시의 LLM 코딩 지적, CLAUDE.md 한 장으로 정리됐다



1 multica-ai 저장소가 안드레이 카파시의 LLM 코딩 지적을 바탕으로 만든 CLAUDE.md 지침이 최근 깃허브 트렌딩에 올랐어요

2 핵심 CLAUDE.md는 가정 금지·최소 코드·최소 수정·성공 기준 네 원칙으로 구성돼 있어요

3 클로드 코드 플러그인이나 커서 규칙 파일로 설치하면 여러 프로젝트에서 같은 지침을 쓸 수 있어요

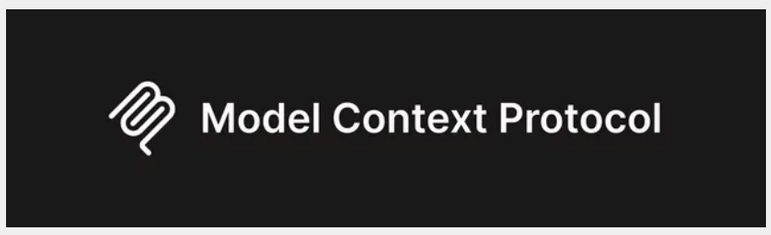
네 번째 원칙은 카파시의 다른 관찰과도 맞닿아 있어요. 목표만 명확히 주면 모델이 스스로 반복하며 그 목표에 맞춰간다는 거고, "make it work"처럼 느슨한 기준을 주면 계속 되물어야 한다는 지적이에요.

multica-ai/andrej-karpathy-skills 저장소는 두 가지 설치 방법을 안내해요. 권장 방법은 클로드 코드 플러그인이에요. 클로드 코드 안에서 먼저 마켓플레이스를 추가하면 가이드라인이 플러그인 형태로 설치되고, 특정 프로젝트가 아니라 클로드 코드로 여는 모든 프로젝트에서 이 스킬을 쓸 수 있게 돼요. 두 번째는 커서(Cursor)예요. 저장소에는 .cursor/rules/karpathy-guidelines.mdc 규칙 파일이 함께 들어 있어서, 커서에서 프로젝트를 열면 같은 원칙이 그대로 적용돼요. 커서용 설정 방법은 저장소 안 CURSOR.md 문서에 따로 정리돼 있어요.

기존에 CLAUDE.md 파일을 쓰고 있는 프로젝트라면 이 네 원칙을 그 파일에 합쳐 쓰면 되고, 프로젝트별로 지켜야 할 규칙이 따로 있으면 지침 아래에 섹션을 만들어 덧붙이면 돼요. 저장소는 이 가이드라인이 오탈자 수정 같은 사소한 작업까지 전부 무겁게 만들려는 게 아니라, 되돌리기 힘든 실수를 줄이는 데 초점을 맞췄다고 설명해요.

◀ 1면에서 계속

MCP 새 로드맵, 서버 상태 없이 수평 확장 지원



1 MCP 코어 메인테이너들이 2026년 3월 로드맵 이후 5개월 만에 새 로드맵을 공개했어요.

2 2026-07-28 스펙 릴리스에서 프로토콜 세션과 초기화 핸드셰이크가 사라져 서버가 상태 없이 수평 확장할 수 있게 됐어요.

3 새 로드맵은 기존 4개에서 5개로 늘어난 우선순위를 제시했고, 에이전트 정체성 인증이 새로 핵심 과제에 올랐어요.

세션 제거 외에도 클라이언트가 서버에 연결하기 전에 server/discover를 호출해 지원 버전과 기능을 먼저 파악할 수 있게 됐고, 목록 조회 결과도 캐싱이 가능해졌어요(SEP-2549). 에이전트 통신 쪽에서는 초기 도입자들의 피드백을 반영해 Tasks 기능을 공식 확장(SEP-2663)으로 옮겼고, 서버가 먼저 요청을 보내던 방식 대신 상태 없는 서버에서도 작동하는 Multi Round-Trip Requests 패턴(SEP-2322)이 새로 들어왔어요.

거버넌스 쪽에서는 Contributor Ladder(기여자 등급 체계)를 공식 채택했고, 각 워킹그룹이 자기 영역의 SEP(스펙 개선 제안)를 직접 심사하도록 바뀌었어요. 스펙에도 정식 기능 수명주기와 폐기 정책이 생겼고, 2026-07-28의 기능 폐기가 이 정책을 처음 따른 사례라고 해요.

엔터프라이즈 준비 작업은 지난 주기 동안 보안에 집중됐어요. 발급자 검증, 발급자에 묶인 클라이언트 자격 증명, 그리고 클라이언트 등록 방식으로 Client ID Metadata Documents(CIMD)를 선호 경로로 도입했고요. 확장 형태로 제공되던 Enterprise-Managed Authorization도 이번에 안정 버전이 됐어요.

◀ 1면에서 계속

크리에이티브

힉스필드, 결혼식 선물 AI 러브스토리로 5억뷰 찍고 프롬프트 공개



1 힉스필드가 지인 결혼식용으로 제작한 AI 러브스토리 단편이 소셜미디어 합산 약 5억 조회수를 기록했다고 밝혔다

2 여섯 시대·두 명의 실존 인물을 배경으로, 클로드와 Seedance 2.5를 힉스필드 플랫폼에서 함께 썼다고 설명했다

3 실사 얼굴 합성, 캐릭터 크기 고정, 손그림 스케치로 공간감 보정 등 다섯 가지 제작 기법과 프롬프트 전체를 무료 공개했다

힉스필드가 이번에 정리한 기법은 실무에서 반복적으로 부딪히는 다섯 가지 문제를 겨냥한다. 첫째는 실사 인물의 얼굴을 AI 영상에 넣을 때 흔히 생기는 '닮은꼴 붕괴' 문제로, 하이브리드 얼굴 합성 방식으로 이를 피했다고 설명했다. 둘째는 캐릭터 크기 문제다. 거인처럼 큰 인물이 장면이 바뀔 때마다 다시 보통 사람 크기로 돌아가 버리는 오류를 캐릭터 스케일을 고정하는 방식으로 잡았다.

셋째는 공간 논리다. AI 모델이 계속 깨뜨리는 장면 속 위치 관계를 연필 드로잉처럼 만든 스케일 스케치를 참조 자료로 넣어 바로잡았다고 밝혔다. 넷째는 군중 장면에서 배경 인물들이 뭉개지는 문제로, 배경 엑스트라를 별도로 캐스팅하는 방식을 썼다. 다섯째는 오디오다. 캐릭터가 노래하는 장면에서 음정이 어긋나지 않도록 오디오 레퍼런스를 붙이는 방법을 공개했다.

원문 튜토리얼이 보여주는 제작 순서는 첫 장면부터 생성하는 방식이 아니다. 먼저 Seedream 5.0 Pro로 캐릭터 시트를 만들고, Soul Cinema로 장소 레퍼런스를 확보한 다음, GPT Image 2로 'The Scale Sketch'를 만들었다. 전체 장면은 바이트댄스 Seed 팀의 Seedance 2.5로 생성했다. 캐릭터와 장소를 먼저 고정해두고 그 위에 장면을 쌓는 방식이다.

RayNeo, 카메라 뺀 AI 안경 'iO 글래스' 공개

스피커도 카메라도 없이 텍스트만 띄우는 웨이브가이드 안경, 구독하면 챗GPT·제미나이도 얹는다



1 RayNeo가 카메라와 스피커를 모두 뺀 AI 안경 iO Glasses를 내놨다

2 초록색 웨이브가이드 디스플레이로 텍스트를 시야에 띄우고 4개 마이크와 골전도 센서로 대화를 인식한다

3 기본은 RayNeo AI와 Gemini 3.1 Flash Lite, 월 9.99달러 구독을 더하면 ChatGPT 5.1·Gemini 2.5 Pro·Claude·DeepSeek을 쓸 수 있다

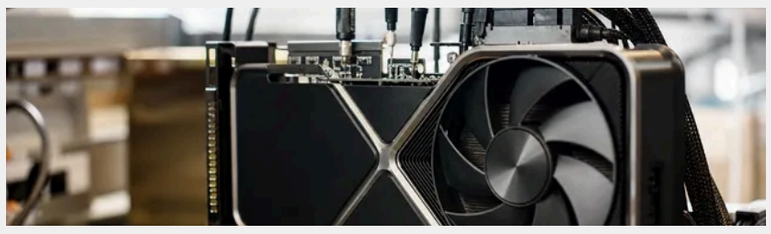
RayNeo iO Glasses는 안경다리나 렌즈 어디에도 카메라가 없다. 스피커도 없다. 대신 초록색 웨이브가이드 디스플레이가 착용자의 시야 안에 텍스트를 띄운다. 투과율은 97%로 렌즈 너머 풍경을 거의 그대로 볼 수 있고, 밝기는 약 1,300니트다. 스마트 안경 하면 흔히 떠올리는 촬영·녹화 기능 대신, 눈앞에 글자를 얹는 데만 집중한 제품이다.

iO Glasses는 대화를 실시간으로 듣고 내용을 요약한 뒤 실행 항목을 뽑아낸다. 회의나 미팅 중 나온 약속을 캘린더 일정으로 제안하고, 착용자가 고개를 끄덕이면 그대로 등록된다. 별도 버튼을 누르거나 화면을 터치할 필요 없이 몸짓 하나로 확인하는 구조다. 프레젠테이션용 텔레프롬프터 모드도 들어 있어 발표 원고를 시야에 띄워 놓고 읽을 수 있다.

음성 인식은 마이크 4개와 골전도 센서가 맡는다. 골전도 센서는 뼈를 타고 전달되는 진동으로 목소리를 잡아내는 방식이라 주변이 시끄러운 카페나 거리에서도 착용자의 말을 구분해낸다. 마이크가 켜져 있을 때는 LED가 점등돼 지금 소리를 듣고 있는 상태인지 겉에서도 확인할 수 있다. 음성-텍스트 번역은 40개 언어를 지원한다.

엔비디아, 하네스만 바꿔 AI 에이전트 점수 30%→100%로

클로드 오퍼스 5에 커스텀 하네스 씌우자 ARC-AGI-3 만점…모델보다 뼈대가 승부처였다



1 엔비디아 연구팀이 클로드 오퍼스 5에 자체 개발한 하네스 AVO를 씌워 ARC-AGI-3 벤치마크에서 100% 점수를 냈다

2 같은 모델을 하네스 없이 돌리면 점수가 30%로 떨어졌는데, 이마저도 테스트한 모델 중 최고치였다

3 오픈AI도 앞서 같은 벤치마크에서 하네스 설정 2개만 바꿔 점수를 3배로 올렸지만 100%에는 못 미쳤다

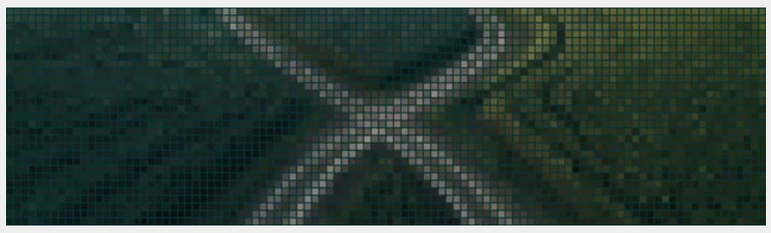
같은 AI 모델을 켜놓고 껍질만 바꿨더니 점수가 30점에서 100점으로 뛰었다. 엔비디아 연구팀이 지난 21일(현지시간) 공개한 실험 결과다. 앤스로픽의 클로드 오퍼스 5를 그대로 썼을 때는 인터랙티브 추론 벤치마크 ARC-AGI-3에서 30%를 기록했는데, 이는 테스트한 모델 중 가장 높은 점수였다. 그런데 이 모델에 엔비디아가 자체 개발한 소프트웨어 껍질을 씌우자 점수가 100%로 올라갔다. 지시문 없이 규칙을 스스로 파악해야 하는 2D 게임 세트인 ARC-AGI-3를 사람처럼 전부 깨버린 셈이다.

하네스는 AI 모델을 감싸는 소프트웨어 뼈대다. 모델에게 어떤 도구를 쓸 수 있게 할지, 이전 대화나 작업 내용을 어떻게 기억하게 할지, 막힌 상황에서 어떤 규칙으로 다음 행동을 정할지를 정해준다. 원문에 따르면 모델 자체는 에이전트의 '두뇌'에 해당하고, 하네스는 그 두뇌가 실제로 손발을 움직여 일하게 만드는 골격이다. 특히 하루 이틀에 끝나지 않고 여러 판단을 며칠에 걸쳐 이어가야 하는 장기 과제(long-horizon task)에서 이 골격의 역할이 두드러진다. 엔비디아가 공개한 연구는 이런 장기 과제에서 모델 선택보다 하네스 설계가 결과를 더 크게 좌우한다고 밝혔다.

엔비디아가 만든 하네스의 이름은 에이전틱 배리에이션 오퍼레이터(Agentic Variation Operators), 줄여서 AVO다. 메모리 관리를 정교하게 조정한 데다, 에이전트가 막다른 길을 계속 헤매거나 이미 지나간 경로를 되풀이할 때 방향을 바로잡아주는 '감독' 역할의 컴포넌트를 추가로 얹었다. 엘 할라크 부사장은 테크크런치에 "감독 에이전트는 마치 CEO처럼 에이전트를 슬쩍 찔러준다"고 설명했다. 감독 컴포넌트 자체가 새로운 개념은 아니지만, 지금 개발자 대부분이 쓰는 클로드 코드나 코덱스, 헤르메스 같은 하네스는 대개 이 감독 계층 없이 단일 계층으로만 돌아간다는 게 엔비디아 쪽 설명이다. 참고로 AVO는 엔비디아가 새로 내놓는 상용 제품이 아니다. 엔비디아는 Nemo 브랜드로 하네스를 만드는 데 필요한 여러 부품을 공개하고 있으며, 일부는 상업용이고 일부는 오픈소스로 풀려 있다.

조지아텍, 올모3 완전공개로 사회적 추론의 뿌리 추적

학습 데이터 12억 문서 중 568만개를 표본삼아 어떤 글이 어떤 능력을 만드는지 확인했다



1 조지아공대 연구팀이 앨런AI연구소의 완전공개 모델 올모3과 학습데이터 돌마3을 이용해 벤치마크별 성능이 어떤 유형의 학습 문서에서 비롯됐는지 추적했다

2 SocialIQA(사회적 추론) 성적은 문학·일상 대화 같은 서사형 글에, ARC와 MMLU는 기술문서·과학 글 같은 설명형 글에 더 크게 의존한 것으로 나타났다

3 연구팀이 사회적 추론과 가장 연관성이 컸던 문학 범주 문서를 모델에서 지우자 무작위로 지웠을 때보다 SocialIQA 점수가 더 크게 떨어져 인과관계를 확인했다

사람의 감정과 의도, 일상적인 도덕적 선택을 읽어내는 능력—이른바 사회적 추론—은 언어모델이 어떤 글을 읽고 배웠는지에 따라 달라진다는 게 조지아공대 연구팀의 결론이다. 연구팀은 모델이 학습한 문서 중 문학 범주에서 영향력이 가장 컸던 문서들을 골라 모델에서 지워봤다. 그러자 사회적 상황을 이해하는 능력을 재는 벤치마크 SocialIQA 점수가, 같은 양을 무작위로 지웠을 때보다 더 크게 떨어졌다. 어떤 글을 배웠는지가 실제로 그 능력을 만든다는 걸 지운 다음 확인한 셈이다.

이 실험을 이끈 건 조지아공대 컴퓨터공학 박사과정의 Glenn Matlin과 공동 저자 Chandreyi(Zini) Chakraborty다. 두 사람은 언어모델의 사회적 추론이 어디서 비롯되는지를 여러 해 추적해왔다. Chakraborty는 설계한 연구에 대해 "질문은 하나였다, [그 능력이] 어디서 오는가"라고 말했다.

연구팀이 쓴 방법은 영향함수(influence functions)라는 기법이다. 모델이 학습한 개별 문서 하나하나가 특정 벤치마크 문제의 답변에 얼마나 영향을 줬는지를 추정하는 방식이다. 이 추정치를 문서 유형별로 합산하면 문학이든 기술 문서든, 어떤 종류의 글이 모델의 특정 능력을 뒷받침하는지 보인다. 다만 이 방법이 성립하려면 모델이 그 문서를 실제로 학습했다는 사실이 확인돼야 한다.

오늘 나온 어려운 말 풀이 전문은 metallab.ai/glossary

소셜아이큐에이
SocialIQA
AI가 사람의 감정·의도·일상적 사회 상황을 얼마나 잘 이해하는지 객관식으로 시험하는 벤치마크

올모3
Olmo 3
앨런AI연구소가 만든 언어모델로, 완성된 결과물뿐 아니라 학습에 쓴 데이터와 중간 과정까지 모두 공개했다.

ARC-AGI-3
ARC-AGI-3
AI에게 규칙을 알려주지 않고 2D 게임을 스스로 풀게 해 얼마나 오래, 스스로 판단하며 문제를 해결하는지 재는 시험이다.

돌마3
Dolma 3
앨런 AI 연구소가 올모3 모델을 학습시키는 데 쓴, 약 12억 6000만개 문서로 이뤄진 공개 학습 데이터 뭉치

오늘 해볼 것

오늘 지면에 나온 것 중 지금 바로 해볼 수 있는 6가지

1 분류기

Classifier

입력된 내용을 미리 정해둔 몇 가지 범주 중 하나로 자동으로 나누는 AI 프로그램.

이게 뭔데

분류기는 들어온 것을 몇 개의 상자 중 하나에 넣어주는 자동 분류원이다. 우체국 직원이 편지 봉투를 보고 '이건 국내 우편함', '이건 국제 우편함'으로 척척 나누듯이, 분류기는 글이나 명령어를 보고 '이건 사람이 쓴 글', '이건 AI가 쓴 글' 또는 '이건 안전한 명령어', '이건 위험한 명령어' 식으로 판정을 내린다.

중요한 건 분류기가 정답을 확실히 안다기보다 확률로 추측한다는 점이다. 수많은 예시를 보고 패턴을 익힌 뒤, 새로운 입력을 봤을 때 '이 패턴이랑 비슷하니 이 상자에 넣자'고 결정하는 방식이라 가끔 오판도 한다. 그래서 분류기 하나만 믿기보다는 사람의 확인이나 다른 안전장치와 함께 쓰이는 경우가 많다.

요즘은 이 분류기가 두 가지 상반된 자리에서 쓰인다. 하나는 AI가 대충 뽑아낸 글을 걸러내는 문지기 역할이고, 다른 하나는 AI 프로그램이 스스로 위험한 명령을 실행하기 전에 브레이크를 거는 역할이다. 둘 다 '이건 통과, 이건 정지'를 판단한다는 점에서 같은 원리다.

해 보려면

챗봇에게 분류기 역할을 직접 시켜볼 수 있어요.

1. 아무 대화형 AI에 다음처럼 요청해요: '아래 문장 다섯 개를 각각 긍정/부정/중립 중 하나로 분류해줘: (문장들을 나열)'
2. AI가 각 문장에 붙인 라벨을 확인해요.
3. 애매한 문장 하나를 골라 '왜 이 라벨을 붙였어?'라고 물어봐요. 분류기가 어떤 단서를 근거로 상자를 골랐는지 짐작해볼 수 있어요.
4. 같은 문장을 표현만 살짝 바꿔서 다시 물어보면, 분류 결과가 달라지기도 해요. 이게 분류기가 확률로 판단한다는 걸 체감하는 방법이에요.

오늘 지면에서 — 링크드인 AI슬롭 신고 버튼, 클릭 100만 건 돌파

2 뮤티카

Multica

안드레이 카파시의 LLM 코딩 지적을 정리해 클로드 코드·커서용 지침 모음으로 공개한 깃허브 개발자(팀)

이게 뭔데

뮤티카는 AI 코딩 도구가 자주 저지르는 실수를 고치기 위한 행동 수칙을 정리해 무료로 나눠준 개발자(팀)예요. 깃허브에서 multica-ai라는 이름으로 활동하며, 안드레이 카파시가 짚은 문제들을 바탕으로 andrej-karpathy-skills라는 지침 모음을 공개했어요.

비유하자면 신입 요리사에게 매번 같은 잔소리를 반복하는 대신, 주방 벽에 붙여둘 규칙표 한 장을 만들어 준 셈이에요. 그 표에는 네 가지 규칙이 담겨 있어요. 멋대로 짐작해서 요리를 진행하지 말고 애매한 부분은 미리 말할 것, 필요한 만큼만 손댈 것, 요청받지 않은 재료는 건드리지 말 것, 그리고 다 만들었다고 끝내지 말고 맛을 확인하는 기준을 정해둘 것이에요.

뮤티카가 만든 건 새로운 AI 모델이 아니라, 기존 AI 코딩 도구에 붙여 쓰는 지침 파일이에요. 이 규칙표를 클로드 코드나 커서 같은 도구에 넣어두면, 도구가 코드를 짤 때마다 그 규칙을 참고하게 되는 방식이에요.

해 보려면

1. 깃허브에서 multica-ai가 공개한 andrej-karpathy-skills 저장소를 찾는다.
2. 클로드 코드를 쓴다면, 안내된 마켓플레이스를 추가해 이 지침을 플러그인 형태로 설치한다. 이렇게 하면 특정 프로젝트뿐 아니라 클로드 코드로 여는 모든 작업에 규칙이 적용된다.
3. 커서를 쓴다면, 저장소 안의 규칙 파일을 프로젝트의 규칙 폴더에 넣어 같은 원칙을 적용한다.
4. 이미 쓰고 있는 지침 파일이 있다면, 새로 받은 네 가지 원칙을 그 안에 합쳐 넣고 프로젝트별로 필요한 규칙은 아래에 덧붙인다.

오늘 지면에서 — 카파시의 LLM 코딩 지적, CLAUDE.md 한 장으로 정리됐다

3 클로드 마크다운 파일

CLAUDE.md

클로드 코드가 작업을 시작하기 전에 항상 읽는, 프로젝트 규칙을 적어두는 메모 파일이다.

이게 뭔데

클로드 마크다운 파일은 클로드 코드에게 일을 시키기 전에 읽히는 사용 설명서 같은 것이다. 새 직원이 출근 첫날 책상에서 발견하는 사내 규칙집을 떠올리면 된다. "이 회사에서는 이렇게 일해요"라고 적어두면, 직원이 매번 물어보지 않고도 그 방식을 따르게 되는 것과 비슷하다.

이 파일이 필요한 이유는 AI 코딩 도구가 자주 저지르는 실수들 때문이다. 사용자에게 확인하지 않고 마음대로 짐작해서 진행하거나, 간단히 끝날 일을 필요 이상으로 복잡하게 만들거나, 요청하지 않은 부분까지 손대는 경우가 있다. 프로젝트 폴더에 클로드 마크다운 파일을 하나 두고 "짐작하지 말고 물어봐라", "필요한 만큼만 코드를 써라", "요청한 부분만 고쳐라" 같은 원칙을 적어두면, 도구가 새로운 대화를 시작할 때마다 이 규칙을 참고해서 움직인다.

특정 프로젝트에만 적용할 수도 있고, 여러 프로젝트에 공통으로 쓰이는 원칙을 담아 공유할 수도 있다. 팀 전체가 같은 방식으로 도구를 쓰게 만드는 손쉬운 방법인 셈이다.

해 보려면

프로젝트 폴더 맨 위에 CLAUDE.md라는 이름의 텍스트 파일을 하나 만들고 아래와 같은 내용을 적어보자.

1. 확실하지 않은 부분이 있으면 짐작해서 진행하지 말고 먼저 물어봐라
2. 문제를 해결하는 데 필요한 최소한의 코드만 작성해라
3. 요청과 관련 없는 코드는 건드리지 마라

실제 인물의 얼굴을 AI 영상에 넣을 때 장면이 바뀌어도 같은 사람처럼 보이도록 여러 참조 이미지를 섞어 얼굴을 고정하는 제작 기법

이게 뭔데

하이브리드 얼굴 합성은 실제 사람의 얼굴을 AI 영상에 넣을 때, 장면이 바뀌어도 계속 같은 사람처럼 보이게 얼굴을 고정하는 방법이다.

화가에게 한 사람의 초상화를 여러 장면에 그려달라고 하면, 매번 조금씩 다르게 그려서 나중엔 다른 사람처럼 보일 수 있다. 하이브리드 얼굴 합성은 화가에게 그 사람의 얼굴 사진 여러 장을 기준 자료로 미리 쥐여주고, 장면을 새로 그릴 때마다 그 사진을 계속 참고하도록 하는 것과 비슷하다. 매 장면마다 인물을 처음부터 새로 설명하는 대신, 미리 정해둔 얼굴 기준을 반복해서 불러오는 셈이다.

이 방법이 필요한 이유는 실사 인물의 얼굴을 AI 영상 생성에 그대로 넣으면 장면이 바뀔 때마다 얼굴이 미묘하게 달라져 다른 사람처럼 보이는 문제가 자주 생기기 때문이다. 하이브리드 얼굴 합성은 이런 문제를 막기 위해 실무에서 쓰이는 제작 노하우에 가깝다.

해 보려면

인물이 등장하는 여러 장면을 만들 계획이라면 다음 순서를 시도해볼 수 있다.

1. 정면, 측면 등 각도가 다른 인물 사진을 여러 장 준비한다.
2. 새 장면을 생성할 때마다 이 사진들을 참조 이미지로 함께 입력한다.
3. 완성된 장면 속 얼굴이 참조 사진과 같은 사람으로 보이는지 장면마다 비교하며 확인한다.

오늘 지면에서 — 힉스필드, 결혼식 선물 AI 러브스토리로 5억뷰 찍고 프롬프트 공개

오늘 나온 말, 조금 더

풀이 전문은 metallab.ai/glossary

소셜아이큐에이

SocialIQA

소셜아이큐에이(SocialIQA)는 인공지능이 사람 사이의 상황을 얼마나 잘 읽어내는지 시험하는 문제집이다. "친구가 약속에 늦어서 화가 났다면 왜 그럴까", "선물을 거절당한 사람은 어떤 기분일까" 같은 짧은 이야기를 주고 여러 답 가운데 가장 그럴듯한 것을 고르게 한다. 사람이라면 굳이 배우지 않아도 아는 것들이지만, AI에게는 감정과 의도, 상식적인 사회적 맥락을 한꺼번에 읽어야 풀리는 까다로운 문제다.

이 시험이 주목받는 이유는 점수 자체보다 "이 능력이 어디서 왔는가"를 캐는 도구로 쓰이기 때문이다. 최근 한 연구는 AI가 학습한 글 가운데 문학 작품을 골라 지워봤더니, 다른 글을 지웠을 때보다 소셜아이큐에이 점수가 유독 크게 떨어진다는 걸 보였다. 어릴 때 소설을 많이 읽은 사람이 남의 마음을 더 잘 헤아리게 되는 것처럼, AI도 대화와 이야기가 많은 글을 배워야 사회적 상황을 잘 읽어낸다는 뜻이다.

올모3

Olmo 3

올모3(Olmo 3)는 앨런AI연구소가 만든 언어모델이다. 보통 '공개된' AI 모델이라고 하면 완성된 요리만 내놓고 레시피는 감추는 식당을 떠올리면 된다. 사용은 할 수 있지만 어떤 재료로, 어떤 순서로 만들었는지는 알 수 없다는 뜻이다. 올모3는 이와 달리 완성품은 물론 재료 목록(학습 데이터인 돌마3), 조리 중간 단계(중간 체크포인트), 맛을 평가하는 기준(평가 도구)까지 전부 열어둔 식당에 가깝다.

이 차이는 단순한 투명성 이상의 의미를 가진다. 모델이 왜 특정 능력을 갖게 됐는지—예를 들어 사람의 감정이나 의도를 읽어내는 능력이 어떤 글을 읽고 배운 결과인지—를 되짚어볼 수 있게 해주기 때문이다. 실제로 조지아공대 연구팀은 올모3의 학습 데이터 중 문학 관련 문서를 골라 지워보는 실험을 통해, 그 능력이 정말로 특정 종류의 글에서 비롯됐는지 확인할 수 있었다. 완성품만 공개된 모델이었다면 애초에 시도할 수 없는 실험이었다.

ARC-AGI-3

ARC-AGI-3

ARC-AGI-3는 AI에게 설명서 없이 2D 게임을 던져주고 혼자 규칙을 알아내 끝까지 깨보라고 시키는 시험이다. 사람이 새로운 오락실 게임을 처음 마주쳤을 때 버튼을 눌러보며 규칙을 스스로 깨우치듯, AI도 시행착오를 거쳐 게임의 논리를 찾아내야 점수를 얻는다.

이 시험이 특별한 이유는 단순히 한 번의 대답을 잘하는지가 아니라, 여러 단계를 거치며 며칠씩 이어지는 긴 판단을 얼마나 흔들림 없이 이어가는지를 본다는 점이다. 그래서 같은 AI 모델이라도 이 모델을 감싸고 있는 소프트웨어 뼈대를 어떻게 짜느냐에 따라 점수가 크게 달라진다. 실제로 어떤 실험에서는 모델은 그대로 두고 뼈대만 바꿨는데 점수가 30점대에서 100점으로 뛴 사례가 나왔다.

즉 ARC-AGI-3는 AI 모델 자체의 똑똑함뿐 아니라, 그 모델을 어떻게 운영하고 관리하느냐가 실제 문제 해결 능력에 얼마나 큰 영향을 미치는지 드러내는 시험대 역할을 한다.

돌마3

Dolma 3

돌마3는 앨런 AI 연구소가 만든 언어모델 올모3를 가르치는 데 쓴 교재 전체라고 보면 된다. 요리에 비유하면, 완성된 요리(모델)만 내놓는 게 아니라 어떤 재료를 얼마나 썼는지 장바구니 영수증까지 통째로 공개한 셈이다. 이 안에는 문학 작품, 일상 대화, 고객상담 글, 기술 문서 등 12억개가 넘는 문서가 카테고리별로 분류돼 담겨 있다.

대부분의 언어모델은 완성품만 공개하고 이 재료 목록은 감춘다. 그래서 어떤 글을 배워서 어떤 능력이 생겼는지 되짚어보기가 사실상 불가능하다. 반면 돌마3는 처음부터 끝까지 열려 있어서, 연구자들이 특정 문서를 빼보고 모델의 능력이 어떻게 바뀌는지 실험할 수 있다. 조지아공대 연구팀이 사람의 감정과 상황을 읽는 능력이 어떤 글에서 비롯되는지 추적할 수 있었던 것도 이 데이터가 통째로 공개돼 있었기 때문이다.

결과적으로 돌마3는 단순한 훈련 재료를 넘어, 언어모델의 능력이 어디서 오는지 확인할 수 있는 실험 도구 역할까지 하고 있다.

텔레프롬프터 모드

teleprompter mode

텔레프롬프터 모드는 발표할 때 대본을 눈앞에 흘려주는 기능이다. 방송 뉴스 앵커가 카메라 옆 유리판에 뜬 원고를 읽으면서도 시청자와 눈을 맞추는 장면을 떠올리면 된다. 스마트 안경에 이 기능이 들어가면, 종이 원고나 손에 든 휴대폰 화면을 내려다볼 필요 없이 청중을 바라본 채로 준비한 말을 읽어나갈 수 있다.

RayNeo의 iO 글래스가 이 기능을 넣은 사례다. 이 안경은 카메라도 스피커도 없고, 렌즈에 초록색 글씨를 띄우는 작은 디스플레