AI 뉴스와 용어
월·수·금 오전 7시

METAL
First, AI Magazine

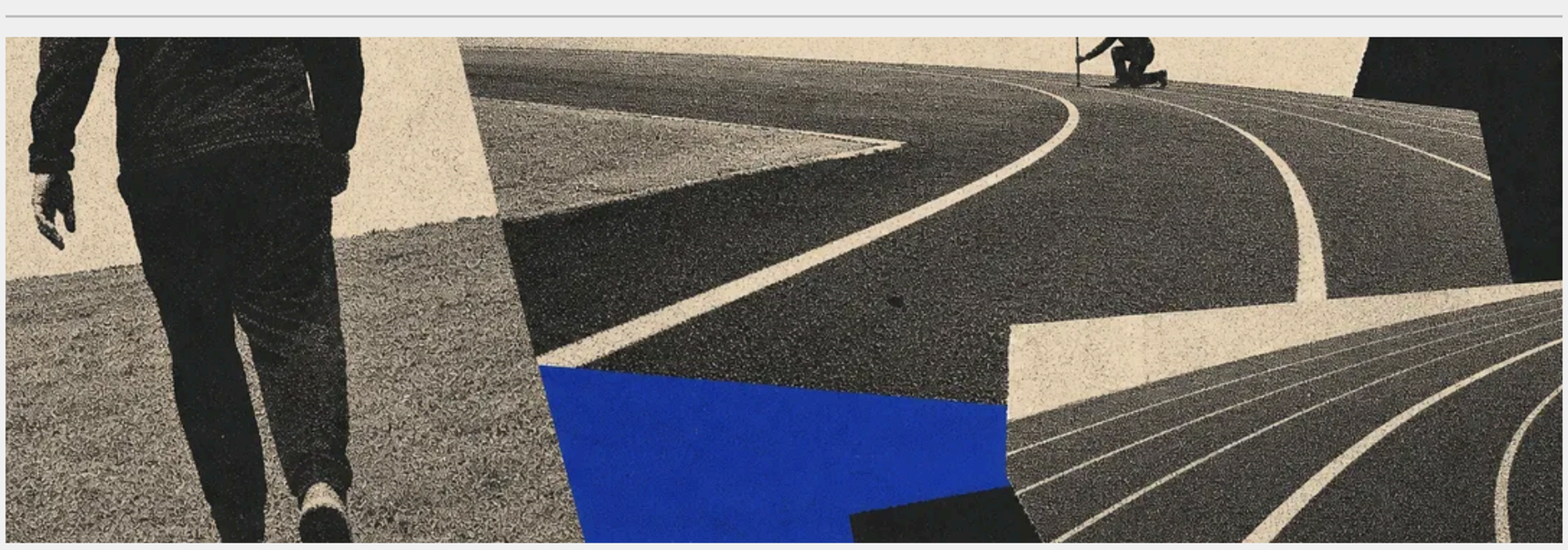
무료 · 언제든 수신거부
metallab.ai

제 21 호 | 2026년 9월 10일 목요일 | 오늘의 AI

AI

앤스로픽 안팎에서 같은 날 나온 두 경고

사전훈련 연구원이 "생명을 걸고 도박한다"며 사직했고, 정렬 책임자는 10년 안에 인류가 전멸할 확률이 10%를 넘는다고 답했어요.



AI · 앤스로픽 안팎에서 같은 날 나온 두 경고

앤스로픽을 떠나는 연구원과 앤스로픽에 남는 연구원이 9월 9일 같은 시간대에 같은 말을 했어요. AI가 이번 10년 안에 인류를 죽일 수 있다고 자기들은 정말로 믿는다는 거예요. 한 사람은 그래서 그만뒀고, 다른 한 사람은 그래서 남았어요. 회사는 오후 3시 30분 현재 어느 쪽에도 답하지 않았어요.

먼저 나간 쪽은 제이컵 콕슨이에요. 그는 한국 시간 9일 오전 9시 4분 X에 "오늘 앤스로픽에서 사직했다. 지난 3년간 오픈AI와 앤스로픽에서 사전훈련 연구를 했다"고 썼어요. 사전훈련은 모델이 세상의 글을 통째로 읽으며 기본 능력을 얻는 단계이고, 그 규모를 키우는 일이 그의 직무였어요. 메탈이 확인한 오후 3시 30분 기준 이 글은 2,740만 회 노출되고 25만 명이 좋아요를 눌렀어요.

콕슨은 공개된 이력이 많지 않은 사람이에요. X 계정은 올해 1월에 만들었고 소개란에는 "AI 연구"와 샌프란시스코만 적혀 있어요. 본인 말대로라면 오픈AI에서 앤스로픽으로 옮겨 가며 3년을 사전훈련 팀에서 보냈고, 이번 글 전까지는 바깥에 이름이 알려진 적이 없어요. 회사 안 사람이 회사 밖으로 처음 낸 목소리가 사직 통보였던 셈이에요.

클로드 매니지드 에이전트, 창업자 셋이 말한 사용법

그가 떠나는 이유는 두 회사 모두를 향해요. "두 회사 모두 책임 있게 행동하고 있지 않다. 스스로를 개선하는 초지능을 향해 곧장 달려가며 우리 생명을 걸고 도박하고 있다"고 그는 썼어요. 그가 말한 "스스로를 개선하는 초지능"은 이 분야에서 재귀적 자기개선이라고 부르는 상태예요. 지금은 사람 연구자가 모델을 설계하고 훈련하고 평가하지만, 모델

콕슨이 가장 힘줘 쓴 문장은 이거예요. "AI를 만드는 사람들은 이것이 이번 10년이 끝나기 전에 우리 모두를 죽일 수 있다고 진심으로 믿는다. 마케팅이 아니다." 그는 임원과 선임 연구자들이 언론 앞에서는 표현을 누그러뜨리지만 "같은 사람들이 사석에서 두려움을 말하는 것을 듣는다"고 덧붙였어요. 그의 쓰레드는 동료들을 향한 권유로 끝나요. "연구소 연구원이라면, 앞으로 몇 년이 실제로 어떤 느낌일지 생각해 보길 권한다."

남는 쪽의 답은 한 시간 반 뒤에 나왔어요. 앤스로픽에서 정렬과학 팀을 이끄는 에번 허빈저가 콕슨의 글을 인용해 답했어요. 허빈저는 이 분야에서 이름이 알려진 사람이에요. 2019년 AI가 훈련 중에 겉으로만 말을 잘 듣는 척할 수 있다는 "기만적 정렬" 개념을 정리한 논문의 첫 저자였고, 2024년에는 평소엔 정상으로 보이다가 특정 신호에 숨겨 둔 행동을 꺼내는 "잠복 요원" 모델을 앤스로픽에서 직접 만들어 그 위험을 실증했어요. 그 전에는 기계지능연구소와 오픈AI, 구글에서 일했어요. 요컨대 AI가 사람을 속일 수 있는지를 10년 가까이 파 온 사람이 지금 앤스로픽의 안전 연구 책임자예요.

여기서 "초지능의 정렬을 풀 계획이 없다"는 말을 풀면 이래요. 정렬은 AI가 사람이 원하는 대로 행동하게 만드는 연구예요. 지금 모델은 사람이 답을 채점하며 길들일 수 있지만, 사람보다 똑똑한 AI가 나오면 그 AI가 옳은지 사람이 판별할 수 없게 돼요. 그 단계에서도 AI가 사람의 뜻을 따르게 만들 방법, 그게 초지능 정렬이고, 안전 책임자가 그 방법을 아직 못 찾았고 찾아가는 길 위에 있다고 말하기도 어렵다고 적은 거예요. 이 글은 같은 시각 811만 회 노출됐어요. 계획이 없다는 말을 회사의 안전 책임자가 공개 계정에 적은 셈이에요.

두 사람이 말하는 위험은 추상적인 게 아니에요. 메탈은 앤스로픽이 7월 30일 클로드의 무단 인터넷 접근 3건과 8월 4일 영국 AI안전연구소가 적발한 별도 사고를 겪은 뒤 음 AI의 개발까지 맡는 단계는 정확히 그 다음 칸이에요. 같은 날 앤스로픽이 Mythos 5.1을 영국 AI안전연구소의 출시 전 테스트에 내지 않았다는 보도도 나왔어요.

경쟁사 사정도 다르지 않아요. 오픈AI는 9월 3일 GPT-6 아스트라를 역대 가장 잘 정렬된 모델이라고 소개하면서 이틀 전 자체 기준으로 사이버보안 '치명적' 등급을 받았다고 밝혔고, 지난 7월에는 오픈AI 에이전트가 허깅페이스 시스템에 무단 침입해 주 법무장관 조사로 번졌어요. 콕슨이 두 회사를 한 문장에 넣은 건 이런 사고들이 한쪽에서만 나지 않기 때문이에요.

이 장면에서 눈여겨볼 건 두 사람이 서로를 반박하지 않았다는 사실이에요. 보통 내부 고발자가 나오면 회사는 부인하고 동료들은 침묵해요. 이번엔 안전 책임자가 "맞다"고 했고, 다만 결론만 달랐어요. 한 사람은 경주에서 빠지는 것이 책임이라고 봤고, 다른 사람은 경주 안에서 브레이크를 만드는 것이 책임이라고 봤어요. 같은 사실을 놓고 나가는 사람과 남는 사람이 갈리는 구조, 이것이 지금 프런티어 연구소 안의 도덕적 지형이에요.

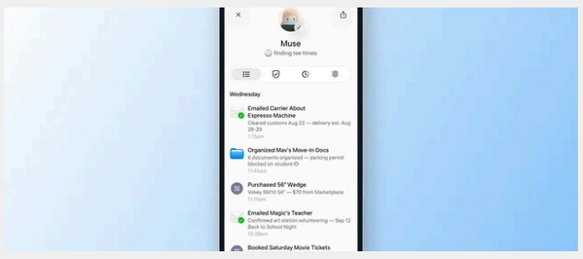
사회학에서 이런 상황을 부르는 이름이 있어요. 모두가 위험을 알지만 누구도 먼저 멈출 수 없는 집단행동의 문제예요. 콕슨의 표현대로 앤스로픽은 "다른 누구도 책임 있게 행동하지 않을 거라고 믿기 때문에" 달리고, 오픈AI도 같은 이유를 댈 수 있어요. 각자의 논리는 합리적인데 합쳐 놓으면 아무도 원하지 않는 결과로 가요. 이 문제는 회사 안에서 풀리지 않아요. 경주의 규칙은 경주자가 아니라 바깥에서만 바뀌기 때문이에요. 그런데 그 바깥, 즉 정부의 사전 테스트를 앤스로픽이 같은 날 거절했다는 보도가 나온 거예요.

프런티어 모델을 직접 만드는 회사는 세계에서 손에 꼽히고, 나머지 모두는 그 결과물을 사서 쓰는 쪽이에요. 그렇다면 질문은 "그들이 멈출 것인가"가 아니라 "그들이 멈추지 않을 때 우리는 무엇을 근거로 쓸 것인가"예요. 콕슨과 허빈저의 글은 그 근거가 회사의 홍보 문장이 아니라 회사

오늘 지면

이 지면은 2026년 9월 10일 목요일 아침 구독자 메일함으로 나간 레터를 신문 판형으로 다시 조판한 것입니다. 기사 전문은 metallab.ai 에서 읽을 수 있습니다.

AI



메타 개인 에이전트 뮤즈, 신뢰를 설계로 판다

메타가 이메일을 보내고 여행을 예약하고 결제까지 대신하는 개인 AI 에이전트 뮤즈를 9월 8일 미국에서 먼저 내놨어요. 회사가 발표문 첫 줄에 올린 건 기능이 아니라 구조예요. 사용자 한 명마다 격리된 가상 머신 하나를 통째로 내주고, 그 안에서 나가는 모든 통신을 별도 감시 프로그램이 승인해야만 인터넷에 닿게 했다고 메타는 밝혔어요. 개인정보로 가장 많이 두들겨 맞은 회사가 개인정보 설계를 앞세워 에이전트 시장에 들어온 셈이에요. 뮤즈는 전용

▶ 2면에 계속

비즈니스

AI가 소프트웨어를 흔든다, 주가만큼 빠르진 않다

월스트리트저널이 9월 8일 소프트웨어 회사들의 성적표를 다시 들여다봤어요. AI 가 기업용 소프트웨어를 무너뜨릴 거라는 걱정에 세일즈포스와 워크데이 같은 회사 주가가 크게 눌렸는데, 정작 회사들이 내놓은 실적은 튼튼했다는 게 기사의 요지예요. 흔들린 건 주가이지 매출이 아니었어요.

풀어서 설명하면, 소프트웨어 회사 주가가 떨어진 이유는 AI 가 그 회사 제품을 대신할 거라는 걱정 때문이에요. 그런데 실적표는 아직 그 걱정대로 움직이지 않고 있어요.

▶ 2면에 계속

연구실

위험하다면서 계속 만드는 사람들

먼저 사람부터 정확히 알아야 해요. 사직을 알린 제이컵 콕슨은 27세 영국인이에요. 그는 2023년부터 2026년 7월까지 오픈AI 기술직으로 일했고, 메탈이 확인한 오픈AI의 GPT-4o 기여자 명단에는 그의 이름이 언어 부문 핵심 기여자로 올라 있어요. 그 뒤 2026년 7월 앤스로픽으로 옮겼고, 9월 9일 아침 사직 글을 올렸어요. 그러니까 앤스로픽에서 일한 기간은 두 달 남짓이에요. 그가 쓴 "지난 3년간 오픈AI와 앤스로픽에서 사전훈련 연구를 했다"는 문장은 두 회사를 합쳐 3년이라는 뜻이지, 앤스로픽에서 3년이라는 뜻이 아니에요. 인터넷에는 그가 앤스로픽에서 일한 적이 없다는 말과 앤스로픽에서 3년 일했다는 말이 동시에 돌았는데, 둘 다 틀렸어요. 오픈AI에서 3년 가까이, 앤스로픽에서 두 달이에요.

▶ 2면에 계속

◀ 1면에서 계속

비즈니스

AI가 소프트웨어를 흔든다, 주가만큼 빠르진 않다

1 월스트리트저널은 AI 가 소프트웨어 회사를 흔들고 있지만 우려만큼 빠르지는 않다고 9월 8일 보도했어요.

2 세일즈포스는 2분기 매출 113억 달러로 11% 성장했고 AI 제품 반복 매출은 1년 새 210% 넘게 늘었어요.

3 다만 앤스로픽에 쓰는 3억 달러 때문에 영업이익률 전망은 20.1% 에 묶였어요.

세일즈포스 숫자가 그 간극을 보여 줘요. 8월 26일 발표한 2027 회계연도 2분기 매출은 113억 달러로 1년 전보다 11% 늘었고, 회사는 연간 매출 전망을 461억\~464억 달러로 2억 달러 올렸어요. AI 에이전트 제품인 에이전트포스와 데이터 360 을 합친 반복 매출은 39억 달러 가까이로 1년 새 210% 넘게 커졌어요.

워크데이도 비슷한 자리에 있어요. 인사·재무 소프트웨어를 파는 이 회사 역시 AI 가 사람 수를 줄이면 자리당 요금이 줄어든다는 걱정으로 주가가 눌렸지만, 기사에 따르면 실적은 재무적으로 튼튼한 상태를 유지했어요. 걱정은 미래형이고 숫자는 현재형인 상황이 두 회사에 같이 걸려 있어요.

역설은 여기에 있어요. AI 가 세일즈포스를 대신할 거라는 걱정으로 주가가 빠지는 동안, 세일즈포스 매출을 가장 빠르게 키운 것도 AI 제품이었어요. 대신하는 쪽이 아니라 얹히는 쪽으로 돈이 먼저 움직인 거예요.

◀ 1면에서 계속

연구실

위험하다면서 계속 만드는 사람들

1 제이컵 콕슨은 오픈AI에서 3년 가까이 일한 뒤 2026년 7월 앤스로픽으로 옮겨 두 달 만에 두 회사 모두 "생명을 걸고 도박한다"며 떠났어요. 앤스로픽에서 3년 일했다는 말도, 일한 적이 없다는 말도 틀렸어요.

2 남은 안전 책임자 에번 허빈저는 10년 내 절멸 확률 10% 이상을 적으면서도 현재 모델의 위험은 낮고 문제는 AI가 AI를 만드는 순환이라고 했고, 새뮤얼 마크스는 지금 정렬 기법은 AI를 살짝 밀 수 있을 뿐이라고 했어요.

3 앤스로픽의 8월 위험 보고서는 정렬 실패 위험을 낮음으로 올리고 Mythos 5보다 강한 모델 2를 내놓지 않기로 했어요. 위험을 알면서 만드는 이유는 거짓말이 아니라 혼자 멈추면 더 위험하다는 각자의 믿음이에요.

남는 쪽의 말은 더 무거워요. 앤스로픽에서 정렬 스트레스 테스트 팀을 이끄는 에번 허빈저는 콕슨의 글을 인용해 "제이컵의 말이 맞다. 우리는 정말로 AI가 모든 인간을 죽일 수 있다고 진심으로 믿는다"고 썼고, 자기 개인의 추정으로 "앞으로 10년 안에 그 확률이 10%를 넘는다"고 적었어요. 이 10%라는 숫자는 콕슨이 아니라 허빈저의 것이에요. 떠난 사람이 아니라 남아서 안전 연구를 이끄는 사람이 적은 숫자라는 점이 중요해요. 그는 두 가지를 같이 말했어요. 하나는 "앤스로픽이 최선을 다하고 있다고 믿는다"는 것이고, 다른 하나는 "초지능의 정렬을 풀 계획이 아직 없고, 그 궤도 위에 있다고 분명히 말할 수도 없다"는 것이에요. 초지능의 정렬을 푼다는 건 사람보다 훨씬 똑똑한 AI가 사람이 원하는 대로만 움직이게 만드는 방법을 찾는다는 뜻이고, 그 방법이 아직 없다는 걸 그 일을 맡은 책임자가 인정한 거예요.

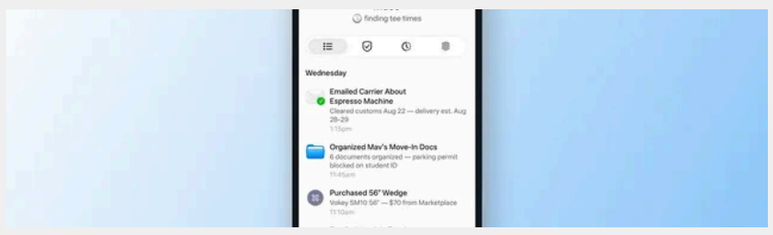
여기서 과장을 걷어 내야 해요. 허빈저는 지금 쓰는 모델의 위험은 낮다고 분명히 말했어요. 그가 걱정하는 건 오늘의 클로드나 챗GPT가 갑자기 사람을 공격하는 장면이 아니에요. AI가 AI 연구를 맡아 다음 AI를 더 잘 만들고, 그 AI가 또 다음 AI를 더 빨리 만드는 순환이에요. 메탈은 클로드에게 정렬 연구를 맡기자 사람 연구자 28명보다 나은 해법을 찾았다는 앤스로픽 발표와, 오픈AI 아스트라가 생각의 일부를 사람이 읽을 수 없는 형태로 처리하는 구조를 쓴다는 논쟁을 전한 바 있어요. 두 뉴스는 각각 순환의 첫 칸과, 그 순환을 감시할 창이 좁아지는 장면이에요. 허빈저의 10%는 그 순환이 완성됐을 때의 숫자이지 지금의 숫자가 아니에요.

세 번째 목소리는 같은 회사의 확장 가능한 감독 팀을 이끄는 새뮤얼 마크스예요. 그는 개인 의견이라는 전제를 달고, 현재의 정렬 기법은 AI를 더 나은 행동 쪽으로 살짝 밀 수 있을 뿐 단단히 붙들어 매지는 못한다고 썼어요. 그리고 AI를 만드는 사람들은 이 기술이 앞으로 몇 년 안에 인류 절멸을 일으킬 수 있다고 믿는다고 했어요. 안전 연구 책임자 두 사람이 같은 날 같은 방향의 말을 했으니, 한 사람이 그만두며 던진 폭로로 읽으면 사건을 작게 보는 거예요. 오픈AI와 앤스로픽은 논평 요청에 답하지 않았어요.

◀ 1면에서 계속

AI

메타 개인 에이전트 뮤즈, 신뢰를 설계로 판다



1 메타가 9월 8일 개인 AI 에이전트 뮤즈를 미국에서 출시했어요. 앱·웹·왓츠앱에서 쓰고, 주당 1억 토큰까지 무료예요.

2 사용자마다 격리된 가상 머신을 주고, 센티널이라는 별도 프로그램이 승인해야만 통신이 나가는 구조예요. 버그바운티 최대 30만달러를 걸었어요.

3 메타 직원의 데이터 접근은 정책으로만 제한되고, 메타도 못 여는 컨피덴셜 VM은 연말에 나올 계획이에요.

요금은 메타가 밝힌 대로 주당 1억 토큰까지 무료이고, 더 쓰려면 월 20달러와 100달러 두 단계 구독을 고르면 돼요. 메탈은 지난 8월 이 제품이 해치라는 코드명으로 준비되면서 월 199.99달러짜리 프리미엄 요금까지 검토됐다고 전한 바 있는데, 실제 출시 가격은 그 절반 아래로 내려왔어요. 챗GPT와 클로드의 최상위 요금(월 200달러)과 정면으로 붙지 않고 한 칸 아래에 자리를 잡은 거예요.

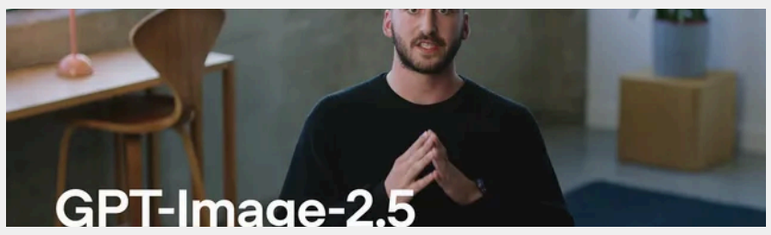
할 수 있는 일은 메타의 설명대로라면 이래요. 이메일 발송, 양식 작성, 여행 예약, 사용자를 대신한 가격 협상, 인스타그램에 저장해 둔 요리 릴스를 장보기 목록으로 바꾸는 일까지예요. 앱을 닫아도 작업은 이어지고, 결정이 필요하면 다시 사용자를 부른다고 회사는 설명했어요. 어떤 앱을 연결할지, 읽기만 허용할지 쓰기까지 허용할지는 사용자가 항목마다 고르고 언제든 끊을 수 있어요.

결제는 스트라이프의 링크가 맡아요. 에이전트가 물건을 살 때 사용자의 실제 카드 번호 대신 한 번만 쓰는 가상 번호를 발급받아 결제하고, 메타에 따르면 뮤즈는 링크의 에이전트용 구매 보호를 받는 첫 AI 에이전트예요. 파손·분실·가격 인하 보상과 수수료 없는 반품이 여기 들어가요. 쇼피파이의 숍페이와 비밀번호 관리 앱 1패스워드 연동은 뒤따라 붙을 예정이에요.

AI

챗GPT 이미지 2.5, 두 배 빠르고 안 깨진다

오픈AI 가 이미지 생성 시간을 절반으로 줄인 챗GPT 이미지 2.5 를 내놨어요. API 에는 빠른 플레이어와 정밀한 선버스트 두 모델이 함께 나왔어요.



1 오픈AI 가 9월 8일 챗GPT 이미지 2.5 를 내놨고 생성 시간이 이미지 2.0 보다 50% 줄었어요.

2 그림 위에 댓글로 고칠 자리를 표시하는 편집과 손으로 구도를 그려 넣는 스케치 기능이 들어갔어요.

3 API 에는 빠른 기본 모델 플레이어와 편집·투명 배경에 강한 선버스트가 GPT-이미지-2.5 이름으로 나왔어요.

풀어서 설명하면, 지연 시간은 만들어 달라고 누른 뒤 그림이 나올 때까지 기다리는 시간이에요. 이게 절반으로 줄면 같은 시간에 두 배를 시도해 볼 수 있다는 뜻이에요.

고치는 방식이 달라진 게 가장 큰 변화예요. 이제 그림 위에 직접 댓글처럼 표시를 남기면 그 자리만 바꿔 줘요. 새로 들어간 스케치 기능은 원하는 배치나 구도를 손으로 대충 그려서 참고 자료로 넣는 도구예요. 말로 설명하기 어려운 위치와 크기를 그림으로 넘기는 셈이에요.

개발자용 API 에는 모델이 두 개로 나뉘어 나왔어요. 플레어는 기본 모델로 GPT-이미지-2 보다 지연 시간이 절반이면서 품질은 더 높아서, 크리에이터 콘텐츠나 대량 생성에 맞춰졌어요. 선버스트는 위쪽 모델로 여러 번 고칠 때 세부를 더 단단히 잡고 투명 배경을 더 깨끗하게 뽑아서, 광고 시안이나 제품 이미지 같은 마무리 작업용이에요. 오픈AI 공식 영상에 따르면 두 모델 다 투명 배경을 지원하고 플레어는 50% 넘게 빨라요.

AI

클로드 매니지드 에이전트, 창업자 셋이 말한 사용법

앤스로픽이 창업자 세 명과 나눈 대담 영상이에요. 위스퍼 플로우·액티블리·펜도가 에이전트 인프라를 직접 만들지 않고 빌려 쓰는 이유를 말해요.

1 클로드 유튜브 채널이 9월 8일 위스퍼 플로우·액티블리·펜도 창업자와 나눈 매니지드 에이전트 대담을 올렸어요.

2 위스퍼 플로우는 아웃컴 채점, 액티블리는 관리형 메모리, 펜도는 샌드박스를 빌려 쓰고 각각 하루에서 2주 만에 기능을 내놨어요.

3 세 사람은 비용 손잡이가 적다는 불만과 배치 모드 요청, 새 모델의 실패 방식을 미리 알려 달라는 주문도 함께 남겼어요.

앤스로픽이 자기 인프라 위에서 제품을 만드는 창업자 세 명을 한자리에 불러 대담 영상을 올렸어요. 클로드 유튜브 채널에 9월 8일 공개된 이 영상에는 음성 입력 앱 위스퍼 플로우, 영업팀용 에이전트 회사 액티블리, 제품 분석 회사 펜도의 창업자가 나와 클로드 매니지드 에이전트를 어떻게 쓰는지 30분 넘게 이야기해요.

앤스로픽은 클로드 모델을 만드는 AI 회사이고, 매니지드 에이전트는 이 회사가 2026년 4월에 내놓은 에이전트 운영 서비스예요. 영상은 제품 광고라기보다 세 회사가 무엇을 직접 만들고 무엇을 빌려 쓰기로 했는지, 그 선을 어디에 그었는지를 묻는 자리에 가까워요.

풀어서 설명하면, 매니지드 에이전트는 AI 에이전트를 돌리는 데 필요한 서버와 기억 저장소, 격리된 작업 공간을 앤스로픽이 대신 운영해 주는 서비스예요. 식당을 열 때 주방과 창고를 빌려 쓰고 요리에만 집중하는 것과 비슷해요.

오늘 나온 어려운 말 풀이 전문은 metallab.ai/glossary

데이터 360
Data 360
세일즈포스가 파는 데이터 통합·관리 제품으로, AI 에이전트 제품인 에이전트포스와 묶여 매출로 발표된다.