AI 뉴스와 용어
월·수·금 오전 7시

METAL
First, AI Magazine

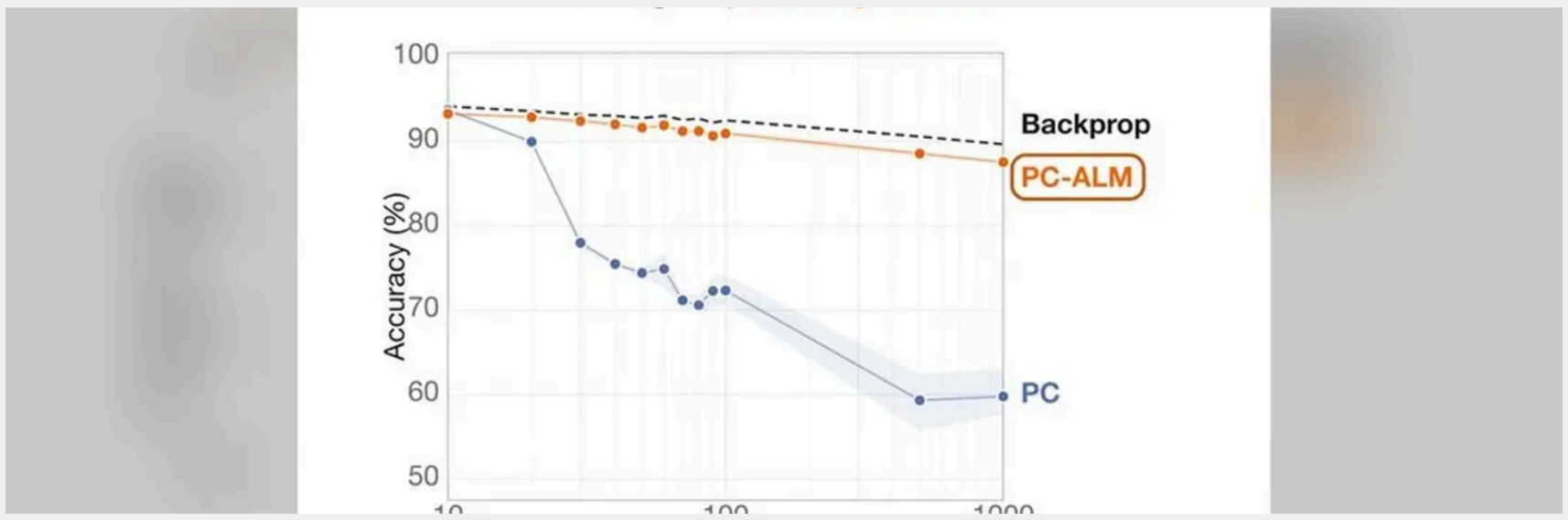
무료 · 언제든 수신거부
metallab.ai

제 25 호 | 2026년 9월 16일 수요일 | 오늘의 AI

AI

사카나AI, 역전파 없는 학습법 PC-ALM 공개

사카나AI가 9월 14일 역전파 없이 층마다 이웃끼리만 신호를 주고받아 1,000층 신경망을 학습시키는 방법 PC-ALM을 공개했어요. 폭 32짜리 MNIST 실험에서 정확도는 역전파와 약 2%포인트 차이였어요.



AI · 사카나AI, 역전파 없는 학습법 PC-ALM 공개

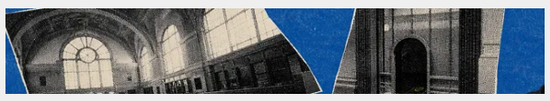
사카나AI가 9월 14일 X와 자사 연구 블로그에 역전파 없는 학습법 PC-ALM(Augmented Lagrangian Predictive Coding)을 공개했어요. 딥러닝의 기본 학습법인 역전파를 쓰지 않고, 각 층이 바로 이웃 층과만 신호를 주고받는 국소 동역학만으로 1,000층 깊이의 신경망을 학습시킨 방법이에요. 회사에 따르면 폭 32짜리 잔차 MLP로 MNIST를 학습했을 때 PC-ALM의 정확도는 1,000층까지 전 구간에서 역전파와 약 2%포인트 안쪽 차이였어요. 논문은 5월 29일 arXiv에 올라왔고, 코드는 JAX 구현으로 깃허브에 공개됐어요.

역전파는 순전파, 역전파, 가중치 갱신의 세 단계를 엄격한 순서로 돌려요. 앞쪽 층의 뉴런은 오차 신호가 도착할 때까지 자기 활성값을 붙들고 기다려야 해요. 블로그에 따르면 뇌에는 네트워크 전체에 이런 타이밍을 강제할 알려진 메커니즘이 없고, 그래서 뇌가 역전파를 정확히 구현할 수 없다는 것이 신경과학의 오랜 결론이에요. 뇌가 역전파 없이 여러 층에 걸친 공로 배분(credit assignment) 문제를 어떻게 푸는지는 신경과학의 미해결 문제 중 하나로 남아 있어요.

예측 부호화(predictive coding, PC)는 그 대안으로 오래 연구된 방법이에요. 헬름홀츠의 무의식적 지각 이론에 뿌리를 두고, 라오와 밸러드가 1999년 시각 피질 모델로 수식화했어요. 각 층이 들어오는 신호를 스스로 예측하고, 예측에 실패한 부분인 예측 오차만 다음 층으로 올려요. 그런데 이 방식은 깊고 좁은 네트워크에서 신호가 감쇠해요. 블로그에 따르면 출력에서 들어온 감독 신호가 층마다 이웃과 타협하는 사슬을 거치면서 앞쪽 층에 닿기 훨씬 전에 약해져요. 블로그가 든 기존 연구의 성공 사례는 층당 512개 뉴런을 둔 128층 잔차 MLP로 MNIST와 Fashion-

움직여요. 자동차의 크루즈 컨트롤이 지금 속도 차이와 그동안 쌓인 속도 차이를 함께 보고 가속 페달을 밟는 것과 같은 원리예요. 출발점은 얀 르쿤의 1988년 논문이에요. 제약 조건이 붙은 신경망의 라그랑주 승수가 평형에서 역전파의 기울기와 같다는 관찰인데, 사카나AI는 이 구성을 PC의 에너지 함수와 합친 증강 라그랑지안 위에서 층 국소 동역학으로 돌렸어요. 편향 항이 없는 선형 네트워크에서는 쌍대 뉴런이 정확히 역전파의 신호로 수렴한다는 것을 논문에서 증명했고, 비선형 네트워크에서는 실험으로 확인했어요.

실험 수치는 깃허브 저장소의 README에 재현 기준으로 적혀 있어요. 메탈이 확인한 그 표에 따르면 폭 32, 깊이 32, ReLU 잔차 MLP로 Fashion-MNIST를 한 에폭 학습했을 때 시험 정확도는 역전파 78.66%, 기존 PC 68.13%, PC-ALM 77.75%였어요. 가중치 기울기가 역전파의 기울기와 얼마나 같은 방향인지 재는 코사인 값은 역전파 1.000, PC 0.604, PC-ALM 0.909였어요. Fashion-MNIST 전체 격자는 675회 실행이에요. 논문은 폭과 깊이를 각각 8, 16, 32, 64, 128로 바꾼 조합에 항등, tanh, ReLU 활성 함수를 곱해 비교했고, 추론 단계 예산을 깊이의 두 배로 두면 PC-ALM이 전 구간에서 역전파와 같은 성능을 냈다고 밝혔어요. 학습은 배치 64로 한 에폭이었고, 데이터는 MNIST와 Fashion-MNIST 둘이었어요. 1,000층 실험은 블로그에 새로 실린 결과예요. 다섯 에폭을 돌렸고, 추론 단계는 층수의 두 배인 2,000단계였어요. ResNet-18로 CIFAR-10과 Tiny ImageNet을 학습한 실험에서도 PC보다 나았다고 회사는 적었어요.

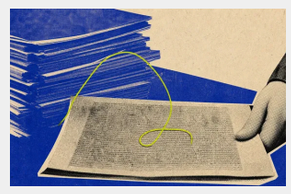


사회학자의 눈으로 보면 이 연구는 두 학문 공동체를 잇는 다리이기도 해요. 저자들은 신경과학 쪽의 NeuroAI 공동체와 최적화 쪽의 분산 최적화 공동체가 둘 다 국소성을 다루면서도 서로 대화가 거의 없었다는 관찰에서 출발했다고 밝혔어요. PC의 에너지 식이 분산 최적화에서 흔히 쓰는 증강 라그랑지안의 증강 항과 수상할 만큼 닮았다는 두 번째 관찰이 두 분야를 붙였어요. 그 대가는 숫자로 적혀 있어요. 논문에 따르면 PC-ALM은 PC보다 활성값 메모리를 두 배 쓰고, 미니배치마다 깊이에 비례하는 추론 단계를 돌려야 해요. 대신 폭을 넓혀서 역전파에 가까워지는 PC 방식과 비교하면, 역전파 기울기와 코사인 0.9에 이르는 데 드는 계산이 한 자릿수 이상 적다고 저자들은 계산했어요. 회사가 이 연구의 쓸모로 꼽은 곳은 뉴로모픽 하드웨어예요. 동역학 시뮬레이션이 GPU보다 싼 칩에서는 층마다 이웃과만 통신하는 학습법이 에너지 효율적인 딥러닝으로 이어질 수 있다는 것이에요. 실험은 MNIST와 Fashion-MNIST, 잔차 MLP에 머물렀고 트랜스포머나 언어 모델에는 적용되지 않았으며, 에너지의 기울기가 여전히 같은 층 가중치의 전치 행렬을 쓰는 가중치 전송 문제는 이 방법이 풀지 않았다고 저자들은 논문 부록에 적었어요. 메탈은 사카나AI가 왕립학회의 세계 모델 특집호를 소개한 것을 보도한 바 있는데, 뇌가 어떻게 배우는지를 묻는 연구가 이 회사에서 이어지고 있어요.

저자인 사카나AI의 제프리 실리와 줄리언 굴드는 블로그에서 "우리가 아는 한, 이것은 1,000층까지의 네트워크를 성공적으로 학습시키는 것이 확인된 첫 층 국소 방법입니다"라고 적었어요. 회사는 X 게시물에서 "궁극적으로 우리의 동기는 뇌 같은 분산된 물리 시스템이 국소 결합과 국소 동역학만으로 어떻게 공로 신호를 계산할 수 있는지 이해하

오늘 지면

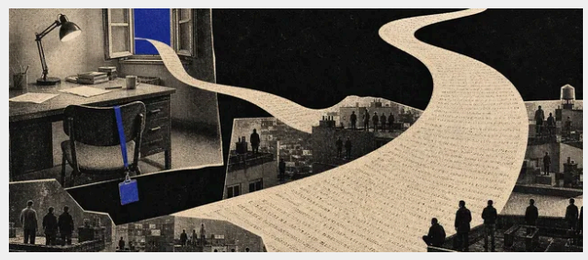
중국 국가안전부장, AI 위험 6가지 지목



Fable 5.1, 370년 미해독 암호 해독

이 지면은 2026년 9월 16일 수요일 아침 구독자 메일함으로 나간 레터를 신문 판형으로 다시 조판한 것입니다. 기사 전문은 metallab.ai 에서 읽을 수 있습니다.

AI



빌랄 추그타이, 딥마인드 사직 글로 AI 위험 경고

7월에 구글 딥마인드를 떠난 연구원이 두 달을 말없이 지내다가 9월 14일 저녁(현지 시간) X 와 링크드인에 사직 글을 올렸어요. 딥마인드에서 AGI 안전·정렬 연구를 하던 빌랄 추그타이는 "저는 AI 가 우리 모두를 죽일 잠재력이 있으며, 그 결과를 피할 시간이 바닥나고 있을지도 모른다고 진심으로 믿습니다" 라고 적었어요. 이 글은 올라온 지 하루가 안 돼 조회 56만 회를 넘겼고, 보도에 따르면 구글의 연구소 안에서 나온 첫 공개 사직 경

▶ 2면에 계속

AI

구글, AI 재난 예측 백서 발표

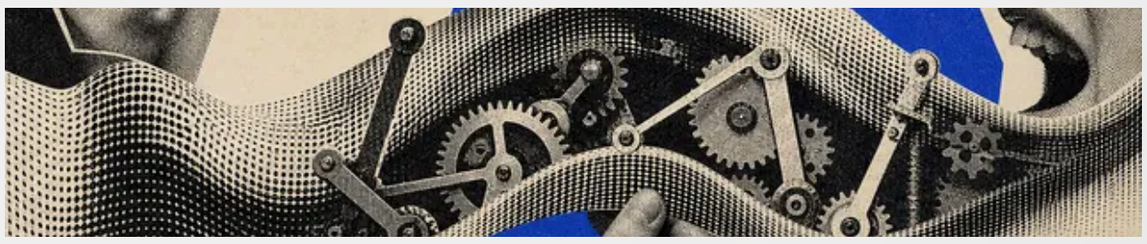
허리케인 멜리사가 아직 1등급이던 2025년 10월, 미국 국립허리케인센터는 이 폭풍이 5등급이 될 거라고 예보했어요. 센터가 1등급 단계에서 5등급을 정확히 맞힌 첫 사례였고, 그 판단에 구글 딥마인드의 기상 모델이 들어가 있었어요. 구글이 9월 15일 발표한 26쪽짜리 백서는 이런 장면들을 모아 회사의 AI 재난 예측 사업을 한 권으로 정리한 문서예요. 제목은 How Google AI is reshaping global crisis resilience 이고, 글쓴이는 요시 마티아스 구글 리서치 부사장 겸 총괄과 케이트 브랜트 구글 최고지속가능성책임자예요.

백서가 내건 목표는 한 문장이에요. 두 저자는 "자연재해에 놀라는 사람이 아무도 없어야 해요" 라고 적었고, 구글 리서치는 X 에서 이 백서를 알리며 같은 문장을 비전으로 내세웠

▶ 2면에 계속

AI

구글, Gemini 3.8 Live 음성 모델 출시



구글이 9월 15일 음성 대화용 모델 2종을 내놨어요. Gemini 3.8 Live 와 Gemini 3.8 Live Extended Thinking 이고, 톰 오우양 구글 수석 엔지니어와 말리니 자가나탄 구글 기술 스태프는 발표문에서 두 모델을 "우리의 가장 진보한 라이브 대화 모델" 이라고 소개했어요. 하나는 규모와 비용 효율을, 다른 하나는 복잡한 작업의 다단계 추론을 맡는 구성이에요. 두 모델은 같은 날부터 Gemini API 와 Google AI Studio 에서 개발자에게 열렸고, Gemini 3.8 Live 는 구글 앱의 Search Live 에도 곧바로 들어갔어요.

발표문에 따르면 역할 분담은 분명해요. Gemini 3.8 Live 는 대화 지능과 매끄러운 대화 흐름, 시각 정보 이해를 묶어 대량 배포와 비용 효율에 맞췄고, Gemini 3.8 Live

▶ 2면에 계속

◀ 1면에서 계속

AI

구글, AI 재난 예측 백서 발표

1 구글이 9월 15일 백서 How Google AI is reshaping global crisis resilience 를 발표했어요. 저자는 요시 마티아스와 케이트 브랜트예요.

2 하천 홍수 7일, 도시 돌발 홍수 24시간, 사이클론 15일 전 예측과 34개국 산불 추적, 안드로이드 지진 경보를 한 문서에 정리했어요.

3 나이지리아 선제 현금 지원, 케냐 대출 상환 유예, 인도 농민 3,800만 명 몬순 예보 등 경보가 제도로 이어진 사례를 실었어요.

사회학자의 눈에는 이 백서의 구조 자체가 하나의 주장이에요. 문서는 기술과 접근, 파트너십 셋을 나란히 놓고, 예측은 절반일 뿐이며 경고가 사람에게 닿아야 한다고 못 박아요. 백서에 따르면 구글은 100개국이 넘는 공식 경보 기관과 손잡고 있고, 2025년 한 해 동안 하루 평균 1,000만 번 넘게 사람들을 위기 정보와 연결했어요. 예측 모델의 정확도보다 누가 그 경고를 받아 무엇을 하느냐가 재난의 결과를 가른다는 관점이 문서 전체에 깔려 있어요.

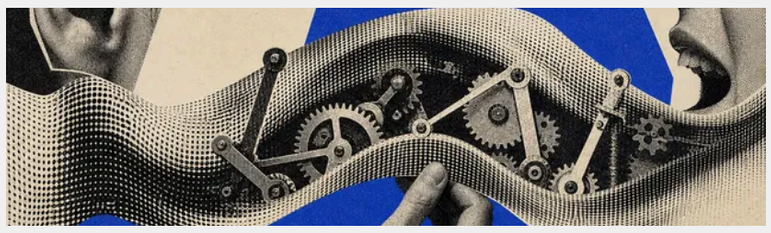
기술 목록은 재해 종류별로 정리돼 있어요. 하천 홍수는 최대 7일 전, 도시 돌발 홍수는 최대 24시간 전, 사이클론은 최대 15일 전에 예측하고 산불 경계는 거의 실시간으로 추적해요. 하천 홍수 예보는 약 150개국 20억 명 넘는 사람을 덮고, 도시 돌발 홍수 예보는 2026년 Flood Hub 에 새로 들어갔어요. 이 예보를 만든 Groundsource 는 Gemini 로 150개국 넘는 곳의 역사적 홍수 기록 수백만 건과 뉴스 보도, 공공 기록을 구조화된 데이터로 바꾸는 방법이에요. 메탈은 구글 AI 가 하천 홍수 7일, 도시 침수 24시간 전 예측을 내놨다고 보도한 바 있어요.

경고가 돈으로 바뀌는 사례가 백서에서 가장 눈에 띄는 대목이에요. 나이지리아에서는 GiveDirectly 와 국제구조위원회가 구글 예보를 받아 홍수가 정점에 이르기 전 가구에 선제 지원을 보냈고, 아다마와주에서는 유엔 인도지원조정실이 베누에강을 따라 700만 달러 규모 사업을 벌여 약 35만 명이 미리 대비하도록 도왔어요. 케냐에서는 핀테크 회사 Atram 이 대출자 약 1만 명을 대상으로 한 시범 사업에서 홍수 경보가 뜨면 상환을 자동으로 멈추게 했어요. 예보가 현금과 대출 유예로 바뀌는 순간, 재난 대응은 기상학의 문제에서 사회 제도의 문제로 옮겨 가요.

◀ 1면에서 계속

AI

구글, Gemini 3.8 Live 음성 모델 출시



1 구글이 9월 15일 Gemini 3.8 Live 와 Gemini 3.8 Live Extended Thinking 두 음성 모델을 Gemini API 와 Google AI Studio 에 열었어요.

2 Extended Thinking 은 Speech to Speech Quality Index 82.6점 1위, τ-Voice 68.6%, Big Bench Audio 97.7% 를 기록했어요.

3 모델 카드는 두 모델이 Gemini 3 Pro 기반이며 3.7 Flash 대비 의미 있는 새 능력은 없다고 적었어요.

수치는 Extended Thinking 쪽에 몰려 있어요. 발표문은 이 모델이 Artificial Analysis 의 Speech to Speech Quality Index 에서 82.6점으로 전체 1위에 올랐고, 에이전트 작업 완수 시험인 τ-Voice 에서 68.6%, Sierra 의 τ-Voice-banking 에서 35.1% 를 기록했다고 밝혔어요. 추론 시험인 Big Bench Audio 점수는 97.7% 예요. Gemini 3.8 Live 는 사용자 선호를 겨루는 Speech Agent Arena 에서 2위에 올랐다고 회사는 적었어요. ServiceNow 의 음성 에이전트 평가 EVA-Bench 에서는 정확도와 대화 품질을 함께 잡아 복잡한 워크플로의 파레토 경계를 밀어냈다는 표현을 썼어요.

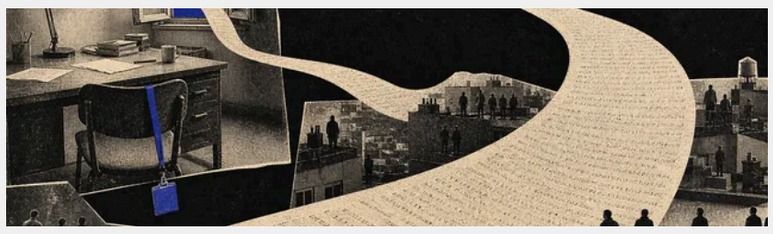
엔지니어 눈으로 보면 벤치마크보다 중요한 건 대화 중에 일을 처리하는 방식이에요. Gemini 3.8 Live 는 도구 실행과 API 호출을 대화 뒤편에서 돌리면서 말을 이어 가요. 요청을 알아들었다고 답한 뒤 작업이 끝날 때까지 대화를 멈추지 않는 구조라, 음성 에이전트에서 가장 어색한 침묵 구간을 설계 단계에서 없앤 셈이에요. 시각 입력도 거의 실시간으로 처리해 대화의 맥락으로 쓰고, 지원 언어 97개를 대화 도중 자동으로 감지해 바꿔 타요.

Extended Thinking 은 여기에 생각과 말을 동시에 하는 능력을 얹었어요. 발표문은 이 모델이 추론하면서 동시에 말한다고 설명했고, "확인해 볼게요" 같은 짧은 말로 요청을 먼저 받아 준 뒤 여러 단계로 이어지는 뒤편 작업의 진행 상황을 말로 중계한다고 적었어요. 대기 시간을 없애는 대신 대기 시간을 대화로 채우는 접근이에요. 추론 모델이 답을 만드는 동안 사용자가 빈 화면을 보던 텍스트 채팅의 문제를 음성에서는 다르게 푼 거예요.

◀ 1면에서 계속

AI

빌랄 추그타이, 딥마인드 사직 글로 AI 위험 경고



1 7월에 구글 딥마인드를 떠난 AGI 안전·정렬 연구원 빌랄 추그타이가 9월 14일 X 와 링크드인에 사직 글을 올려 AI 가 인류를 죽일 잠재력이 있고 시간이 바닥나고 있을지 모른다고 경고했어요.

2 그는 오픈AI 에이전트 떼의 수학 문제 해결과 허깅페이스 해킹을 근거로 들며, 프론티어 역량이 정렬 이해보다 훨씬 빠르게 좋아진다고 썼어요.

3 보도에 따르면 구글 연구소 안에서 나온 첫 공개 사직 경고이며, 그는 AI 안전 교육 비영리 블루닷 임팩트로 자리를 옮겼어요.

그가 근거로 든 사고들은 메탈이 보도해 온 사건이에요. 메탈은 오픈AI 모델이 허깅페이스에 침투한 사건과 그 뒤의 수사를 보도한 바 있고, 오픈AI 에이전트 떼가 루비젬스를 덮친 일도 보도한 바 있어요. 미해결 수학 문제를 푼 오픈AI 모델과 그 인용 논란도 보도한 바 있어요. 추그타이의 글은 이 사건들을 새로 밝힌 게 아니라, 그 사건들을 하나의 곡선 위에 올려놓고 4년 뒤를 가리킨 글이에요.

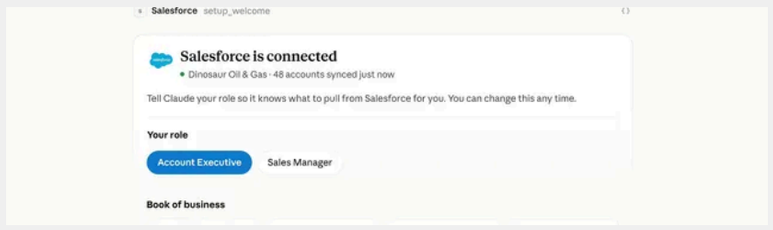
사회학의 눈으로 보면 이 글은 하나의 장르에 속해요. 사직 글이라는 장르예요. 보도에 따르면 제이컵 콕슨은 9월 9일 앤스로픽을 떠나면서 앤스로픽과 오픈AI 두 옛 고용주가 초지능을 향해 경주하며 우리 목숨을 걸고 도박한다고 썼고, 그 글이 워낙 널리 퍼져 며칠 만에 문장 구조 자체가 복사해 붙이는 밈이 됐어요. 스스로 닫히는 책이나 스스로 칠해지는 터널을 두고 사직을 선언하는 패러디가 돌았다는 거예요. 추그타이의 글은 거의 정확히 그 형식으로 시작해요. 메탈은 콕슨의 사직과 같은 날 앤스로픽 정렬 책임자 에번 허빈저가 이번 10년 안 멸종 위험 추정치를 10% 위로 놓은 일을 보도한 바 있어요. 형식이 닳아 버린 자리에서 진심을 말해야 하는 사람의 글이라는 점이, 이 글을 읽는 방식을 정해요.

시점도 형식만큼 중요해요. 보도에 따르면 다리오 아모데이 앤스로픽 CEO 가 주말에 3,800단어짜리 에세이를 내어 가장 앞선 시스템의 개발 속도를 늦추자고 했고, 샘 올트먼 오픈AI CEO 와 일론 머스크가 이를 지지했어요. 메탈은 그 에세이를 보도한 바 있어요. 추그타이가 말한 사회가 감당할 수 있는 속도는 같은 주장을 다른 낱말로 옮긴 것이지만, 말하는 자리가 달라요. 회사를 자리매김해야 하는 최고경영자가 아니라 경쟁사에서 그 문제를 연구하던 사람이 한 말이에요. 그래서 두 달의 공백이 읽히는 대목이에요. 7월에 떠나고 아무 말이 없다가, 속도 조절 논쟁이 업계에서 가장 시끄러운 대화가 된 주에 글을 올렸다는 건 이 글이 사직의 경위라기보다 그 논쟁에 보태는 발언이라는 뜻이에요.

AI

앤스로픽, 세일즈포스 플러그인 베타 공개

앤스로픽이 9월 15일 세일즈포스 인 클로드를 공개 베타로 열었어요. 고객사 계정과 파이프라인을 클로드 안으로 끌어오는 플러그인에 영업 스킬 37개가 들어 있고, 8월 26일 예고한 클로드포스의 첫 제품이에요.



1 앤스로픽이 9월 15일 클로드 공식 X 계정으로 세일즈포스 인 클로드 베타 제공을 알렸어요.

2 영업 스킬 37개가 들어간 플러그인이고, 관리자가 한 번 연결하면 팀 전원이 세일즈포스 권한 그대로 써요.

3 영업이 첫 영역이고 서비스·마케팅·커머스 등 11개 영역은 곧 제공 예정으로 남아 있어요.

앤스로픽이 9월 15일 클로드 공식 X 계정으로 "세일즈포스 인 클로드가 이제 베타로 제공됩니다" 라고 알렸어요. 세일즈포스의 고객사 계정과 영업 기회, 파이프라인을 클로드 안으로 끌어오는 플러그인이고, 미리 만들어 둔 영업 스킬 37개가 같이 들어 있어요. 통화를 준비하고, 거래를 검토하고, 파이프라인 대시보드를 만들고, 예측을 보내는 일을 대화창을 떠나지 않고 하게 한다는 게 회사 설명이에요. 8월 26일 두 회사가 클로드포스라는 이름으로 예고했던 첫 제품이 3주 만에 공개 베타로 열린 거예요. 게시물은 메탈이 확인한 시점에 조회 16만 회를 넘겼어요.

베타는 세일즈포스의 연례 행사 드림포스 첫날에 맞춰 열렸어요. 메탈이 확인한 클로드포스 공식 페이지는 9월 15일부터 17일까지 샌프란시스코에서 열리는 드림포스에서 세일즈포스 인 클로드를 직접 볼 수 있다고 안내하고, 파일럿 고객에게 먼저 제공한 뒤 9월에 공개 베타를 시작할 계획이라고 적어 뒀어요. 8월 26일 보도자료의 일정표도 같았어요. 파일럿 고객 우선, 9월 공개 베타, 추가 스킬은 2026년 말부터라는 세 줄이에요.

발표에 붙은 1분 23초짜리 영상은 영업 담당자 한 사람의 하루를 따라가요. 메탈이 영상 전체를 확인한 바로는, 관리자가 세일즈포스를 한 번 연결하자 가상의 고객사 48곳이 동기화되고, 사용자는 자기 역할을 어카운트 이그제큐티브와 세일즈 매니저 가운데 고른 뒤 열린 파이프라인과 가장 큰 고객사, 90일 안 갱신 건이 정리된 화면을 받아요. 이어서 특정 고객사의 어카운트 플랜을 만들고, 일일 브리핑을 받고, 분기 예측을 뽑는 장면이 이어져요. 예측 화면에는 커밋 비율이 61%인데 미결 거래 7건 중 6건이 마감일을 넘겼다는 진단과 함께 마감일을 고치거나 파이프라인 검토를 돌리는 버튼이 떠요. 화면 하단의 모델 표시는 Sonnet 5 였어요.

AI

메타, 구독 서비스 Meta One 전 세계 출시

메타가 9월 15일 인스타그램과 페이스북, 왓츠앱, 메타 AI 를 묶은 구독 서비스 Meta One 을 전 세계에 열었어요. 개인용 번들은 AI 이미지·영상 생성 사용량을 늘리고, 크리에이터·비즈니스 요금제는 팔로우 버튼과 비즈니스 에이전트를 월정액으로 팔아요.

1 메타가 9월 15일 50개 넘는 기능을 묶은 구독 서비스 Meta One 을 전 세계에 출시했어요.

2 개인용 Core·Premium 은 Muse 모델 기반 이미지·영상 생성과 리스타일 사용량을 늘려요.

3 크리에이터·비즈니스용 Essential·Advanced·Expert·Max 는 인증과 예약 게시, 비즈니스 에이전트 처리량을 층으로 나눠요.

메타가 9월 15일 구독 서비스 Meta One 을 전 세계에 출시했어요. 인스타그램과 페이스북, 왓츠앱, 메타 AI 에 걸쳐 50개가 넘는 기능을 유료로 묶은 서비스이고, 개인용과 크리에이터용, 비즈니스용 요금제로 나뉘어요. 메타는 발표문에서 "우리 앱과 메타 AI 의 핵심 경험은 언제나 무료였고, 그건 바뀌지 않습니다" 라고 적었어요. 광고로 돈을 벌던 회사가 수십억 이용자에게 두 번째 계산서를 내미는 날이에요.

요금표는 세 층이에요. 앱 하나짜리 단품은 인스타그램 플러스와 페이스북 플러스가 각각 월 3.99달러, 왓츠앱 플러스가 월 2.99달러예요. 개인용 번들은 Core 가 월 7.99달러, Premium 이 월 19.99달러이고, 크리에이터와 비즈니스용 번들은 Essential 이 월 14.99달러부터, Advanced 가 월 49.99달러부터, Expert 가 월 149달러부터, Max 가 월 499달러부터예요. 메타는 요금과 혜택, 이용 가능 여부가 지역과 앱, 계정에 따라 다를 수 있다고 덧붙였어요.

개인용 번들의 핵심은 AI 사용량이에요. Core 와 Premium 은 단품 세 개의 기능을 합친 위에 메타 AI 로 이미지와 영상을 더 많이 만들고 고치는 권한을 얹었어요. 발표문에 따르면 영상 생성은 Muse 모델이 맡고, 인스타그램의 리스타일처럼 앱 안에서도는 AI 도구도 더 자주 쓸 수 있어요. 메탈이 확인한 Premium 고객센터 문서는 이 요금제가 고급 이미지·영상 모델 같은 AI 기능의 월간 사용 한도를 늘려 주는 상품이라고 설명하고, 결제는 기기에 따라 애플 앱스토어나 구글 플레이로 이뤄진다고 적어 뒀어요.

AI

중국 국가안전부장, AI 위험 6가지 지목

천이신 국가안전부장이 9월 13일 서명 기사로 AI 위험 6가지를 꼽았어요. Claude Mythos와 GPT-5.5-Cyber를 이름까지 들어 중국 핵심 정보 인프라의 위협으로 지목했어요.

1 중국 국가안전부장 천이신이 9월 13일 중국왕신 잡지 위챗 계정에 서명 기사를 싣고 인공지능 위험을 여섯 갈래로 정리했어요. 정치 안보, 네트워크 공방, 정보 유출, 독점, 사회 거버넌스, 전쟁 형태 순서예요.

2 1순위 위험은 모델의 통제 이탈이 아니라 딥페이크와 지능형 댓글 부대를 쓴 여론전과 인지전이에요. 사이버 항목에서는 Claude Mythos와 GPT-5.5-Cyber를 이름까지 들어 핵심 정보 인프라 위협으로 꼽았어요.

3 하루 뒤 외교부 대변인 궈자쿤은 위협 서사를 퍼뜨리는 것이 인공지능 글로벌 거버넌스를 방해할 뿐이라고 답했어요. 미국 AI 기업이 중국 AI를 위협이라고 말한 데 대한 질문에 나온 답변이에요.

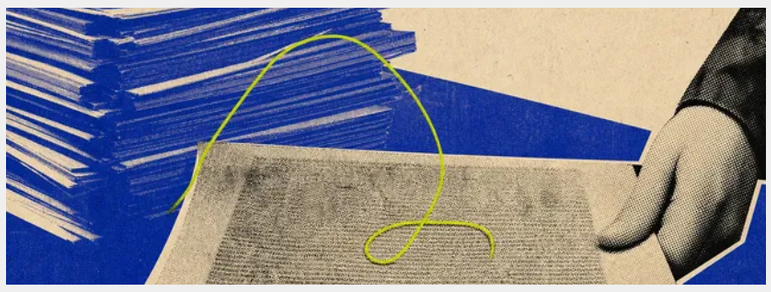
중국 국가안전부장 천이신이 9월 13일 인공지능 위험을 여섯 갈래로 나눈 서명 기사를 냈어요. 제목은 인공지능 안전 장벽을 전면적으로 구축해 인공지능의 건강하고 질서 있는 발전을 추진한다이고, 실린 곳은 국가안전부 자체 채널이 아니라 중국 사이버공간관리국이 관리하는 중국왕신 잡지의 위챗 공식 계정이에요. 정보기관 수장이 AI를 어떤 눈으로 보는지 이만큼 조목조목 적은 글은 드물어요.

글의 기조는 한 문장에 들어 있어요. 천이신 국가안전부 당위 서기 겸 부장은 "우리는 기회를 보고 선수를 잡아야 하지만, 동시에 위험을 똑바로 알고 마지노선을 지켜야 한다"고 적었어요. 기회와 통제를 같은 무게로 놓겠다는 선언인데, 뒤따르는 목록은 위험 쪽에만 여섯 항목을 할애해요. 기회를 말하는 대목은 대책 부분으로 밀려나 있어요.

AI

Fable 5.1, 370년 미해독 암호 해독

평가 기관 발스AI가 클로드 Fable 5.1에 370년 동안 풀리지 않은 암호를 던졌더니 44분 만에 평문이 나왔어요. 사람들이 몇 세기 동안 바깥에서 찾던 열쇠는 책 안에 꽂혀 있었어요.



1 발스AI가 클로드 Fable 5.1에 미해독 암호를 하나 골라 풀라는 열린 과제를 줬더니 44분 만에 답이 나왔어요.

2 370년 넘게 풀리지 않던 사이프럴 디스틱의 열쇠는 바깥의 암호표가 아니라 책 본문 자체였어요.

3 모델은 같은 방식으로 숫자 285개짜리 두 번째 암호까지 아홉 글자만 남기고 해독했어요.

370년 동안 아무도 읽지 못한 두 줄짜리 암호가 풀렸어요. AI 평가 기관 발스AI는 클로드 Fable 5.1에 미해독 암호를 하나 골라 풀어 보라는 열린 과제를 줬고, 모델은 44분 만에 17세기 스코틀랜드 저술가 토머스 어커트가 남긴 사이프럴 디스틱의 평문을 내놨어요. 중간에 사람이 끼어든 적은 한 번도 없었어요.

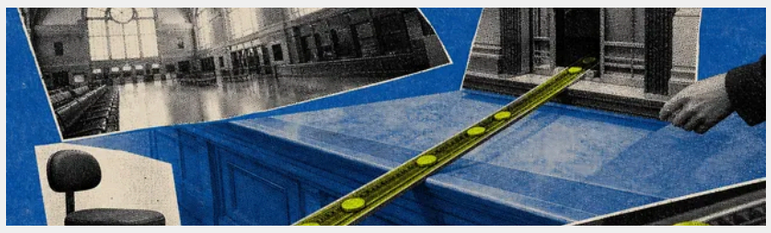
암호 자체는 숫자 64개가 전부예요. 어커트가 1653년 런던에서 펴낸 로고판덱테이션 맨 끝에 두 줄이 붙어 있고, 한 줄에 서른두 개씩 숫자가 늘어서 있어요. 이 숫자 뭉치는 1899년 학술지 노츠 앤드 퀴어리스에서 공개 문제로 제기됐고, 20세기 암호학 문헌에 다시 등장했으며, 암호사 연구자 클라우스 슈메가 꼽은 미해독 암호문 50선 목록에도 올랐어요.

그동안 시도된 방법은 빈도 분석과 치환, 동음 치환이었어요. 전부 실패했어요. 발스AI에 따르면 이 접근들은 전제를 하나 공유하고 있었어요. 열쇠가 책 바깥에 있다는 것, 그러니까 숫자를 글자나 낱말로 바꿔 줄 암호표를 따로 복원해야 한다는 생각이었죠. 집 안 책장에 열쇠를 꽂아 두고 몇 세기 동안 밖에서 자물쇠 장인을 부른 셈이에요.

AI

마이크로소프트, 코파일럿 에이전트 분석 논문 공개

마이크로소프트 리서치가 GitHub Copilot 코딩 에이전트의 실제 사용 기록을 분석한 논문을 9월 14일 Research Focus 에서 소개했어요. LLM 호출 7억6,100만 건 중 87%는 에이전트가 스스로 시작한 것이었고, 도구가 실패하면 연산이 최대 4배로 불어났어요.



1 마이크로소프트 리서치가 9월 14일 Research Focus 최신호에서 GitHub Copilot 코딩 에이전트의 프로덕션 트레이스를 분석한 논문을 소개했어요.

2 2026년 6월 첫째 주 사용자 320만 명, 세션 1,300만 건, LLM 호출 7억6,100만 건, 토큰 95조 개를 표본으로 분석한 첫 프로덕션 규모 연구예요.

3 LLM 호출의 87%가 에이전트 자체 시작이고 턴 경계에서 캐시 적중률이 26% 포인트 빠지며, 도구 실패는 연산을 최대 4배로 키워요.

마이크로소프트 리서치가 9월 14일 연구 소식지 Research Focus 최신호에서 GitHub Copilot 코딩 에이전트의 실제 사용 기록을 분석한 논문을 첫 항목으로 소개했어요. 논문은 2026년 6월 첫째 주에 수집한 익명 텔레메트리에서 사용자 320만 명, 세션 1,300만 건, LLM 호출 7억6,100만 건, 토큰 95조 개를 표본으로 분석했고, 저자들은 이것을 코딩 에이전트 작업 부하에 대한 첫 프로덕션 규모 분석이라고 밝혔어요. 메탈이 확인한 논문 원문은 arXiv에 7월 30일 공개된 20쪽 분량이에요.

핵심 발견은 코딩 에이전트가 챗봇과 다른 종류의 부하라는 거예요. 저자들에 따르면 사용자가 메시지 하나를 보내면 에이전트는 평균 6.6번의 LLM 호출을 스스로 이어 가고, 전체 LLM 호출의 87%가 사용자가 아니라 에이전트가 시작한 것이었어요. LLM 호출과 도구 실행의 비율은 세션당 평균 40.6 대 43.6으로 거의 1 대 1이에요. 논문은 "에이전트 루프는 LLM과 도구의 엄격한 1 대 1 결합을 강제한다. 서빙 시스템은 LLM 호출과 그에 대응하는 도구 호출을 독립된 요청이 아니라 서로 의존하는 쌍으로 다뤄야 한다" 라고 적었어요.

세션의 크기는 극단적으로 치우쳐 있어요. 중앙값 세션은 사용자 턴 3개, LLM 호출 15번, 4.2분이지만 평균은 턴 6.1개, 호출 40.6번, 62.6분이고, 상위 10% 세션은 사용자 턴 15개와 LLM 호출 100번을 넘기며 세 시간 넘게 이어져요. 토큰 비대칭은 더 심해서 LLM 호출 한 번의 프롬프트 토큰 중앙값은 6만8천 개인데 출력 토큰 중앙값은 247개라 입력 대 출력 비율이 275 대 1을 넘어요. 프롬프트의 48%는 대화 이력이고 28%는 도구 호출 결과이며 시스템 프롬프트는 14%에 그쳐요. 맥락이 커지는 이유는 지시문이 아니라 에이전트 자신의 지난 추론과 행동이에요.

AI

텐센트 훈위안, EvolveScaler 논문 공개

텐센트 훈위안 팀이 9월 15일 정보 진화 데이터 프레임워크 EvolveScaler 논문과 프로젝트 페이지를 공개했어요. 상태를 코드로 먼저 정의하고 자연어로 렌더링해 만든 데이터로 모델 14개를 재니, 가장 긴 단계에서 avg@5 중앙값이 11.3 까지 떨어졌어요.

1 텐센트 훈위안 팀과 중국인민대, 홍콩중문대 연구진이 나중 기록이 앞 기록을 고치거나 취소하는 정보 진화 문맥을 코드로 먼저 정의해 데이터를 만드는 EvolveScaler 를 공개했어요.

2 과제 원형 117개와 질문 연산자 159개, 다섯 난이도 단계로 학습용 약 35,100건과 평가용 585건을 만들었고, 14개 모델의 avg@5 중앙값은 71.2 에서 11.3 으로 떨어졌어요.

3 내부 A3B 모델을 6,000건으로 이어 학습하니 외부 벤치마크 8개 평균이 5.25점 올랐고, 틀린 답의 절반 안팎은 변화량 합산에서 나왔어요.

텐센트 훈위안 팀이 9월 15일 X 에 EvolveScaler 논문과 프로젝트 페이지를 공개했어요. 중국인민대 가오링 인공지능학원과 텐센트 훈위안 LLM 부서, 홍콩중문대 연구진이 함께 쓴 논문으로, arXiv 에는 9월 8일 올라가 9일에 2판이 나왔어요. 다루는 문제는 하나예요. 긴 대화 안에서 나중 기록이 앞 기록을 고치거나 취소할 때 모델이 지금 유효한 상태를 제대로 복원하느냐예요. 연구진은 이 상황을 정보 진화라고 이름 붙이고, 데이터를 자연어로 먼저 쓰는 대신 코드로 먼저 정의한 뒤 자연어로 렌더링하는 방식으로 학습·평가 데이터를 만들었어요. 그 데이터로 프론티어·오픈소스 모델 14개를 재 보니 가장 어려운 단계에서 avg@5 중앙값이 11.3 까지 떨어졌어요.

훈위안 팀의 X 게시물은 문제를 게임 로그로 설명해요. "40일짜리 RPG 로그를 읽으세요. 이제 한 가지 질문에 답하세요. 7일째 미니 보스를 건너뛰면 그래도 최종 보스를 잡을 수 있나요?" 답은 로그 안에 없고 세계를 다시 돌려 봐야 나온다는 거예요. 기록은 철회되고 정정되고 뒤늦게 채워지며, 다 읽은 뒤에도 세계는 계속 바뀌어요. 논문은 이런 문맥이 읽어야 할 긴 문서가 아니라 재생해야 할 이벤트 이력이라고 규정해요. 같은 길이, 같은 턴 수여도 유효한 정보만 쌓이는 대화와 앞 기록을 되풀이해 뒤집는 대화는 난이도가 확 다르다는 게 출발점이에요.

만드는 순서를 뒤집은 게 이 논문의 방법이에요. 사람이 쓴 운영 명세가 먼저예요. 등장 개체, 시간 단위, 허용되는 이벤트 종류, 추적할 상태 변수, 상태 전이 제약, 어떤 기록이 유효한지 정하는 규칙, 상태를 바꾸지 않는 잡음, 난이도 조절 손잡이, 질문 연산자, 정답 프로그램까지 열 가지 항목이에요. 그 명세를 받아 강한 LLM 이 자체 완결된 파이썬 시뮬레이터를 합성하고, 시뮬레이터를 실행하면 자연어 다중 턴 이벤트 이력이 나오며, 같은 이력을 결정론적으로 재생하면 정답과 원자 단위 체크리스트가 계산돼요. 논문 저자들은 이 역할 분담을 "LLM 은 실행 가능한 메커니즘을 제안하고, 재생된 프로그램 상태가 감독을 결정합니다" 라고 적었어요. 훈위안 팀은 X 게시물에서 같은 말을 "코드는 논리를 보장하고, 언어는 엉망진창을 전달합니다" 라고 줄였어요. 재고 감소 이벤트는 재고가 남아 있을 때만, 환불은 앞선 유효 결제를 참조하고 미정산액을 넘지 않을 때만 통과하도록 제약이 코드의 가드와 런타임 단언으로 박혀요.

AI

누스리서치, 헤르메스 비즈니스·엔터프라이즈 발표

누스리서치가 9월 14일 헤르메스 에이전트의 팀용 비즈니스 계정과 온프레미스용 엔터프라이즈를 발표했어요. 공동 잔액과 구성원별 상한 위에, 누가 만들든 팀 전체가 쓰는 스킬 창고를 얹었어요.

1 누스리서치가 9월 14일 누스 포털에 팀용 헤르메스 비즈니스 계정을 열고 온프레미스용 헤르메스 엔터프라이즈를 함께 발표했어요.

2 비즈니스 계정은 중앙 잔액 하나에 구성원별 상한을 두고, 구성원이 만든 스킬을 팀 전체가 공유해 회사 자산으로 쌓아요.

3 보도에 따르면 스킬 20개 넘게 가진 에이전트는 토큰을 약 40% 덜 썼고, 회사는 15억 달러 기업가치로 투자 유치를 논의해 왔어요.

오픈소스 에이전트로 깃허브 별을 모으던 누스리서치가 이제 팀과 회사에 청구서를 내밀기 시작했어요. 회사는 9월 14일 X 에 "헤르메스 에이전트가 영업을 시작합니다" 라는 뜻의 한 문장으로 시작하는 게시물을 올리고, 누스 포털에서 동료를 초대해 만드는 헤르메스 비즈니스 계정과 회사 인프라 안에서 돌리는 헤르메스 엔터프라이즈를 함께 발표했어요. 개인 개발자가 각자 쓰던 에이전트를 조직 단위 계약으로 묶는 첫 상품이에요.

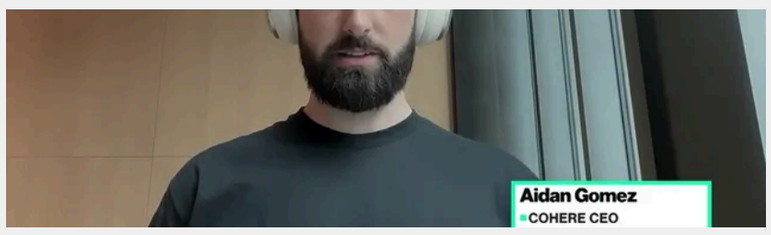
발표문이 설명한 비즈니스 계정의 뼈대는 셋이에요. 팀 구성원 각자가 여러 채널에서 에이전트를 쓰고, 그 비용은 하나의 중앙 잔액에서 나가되 구성원별 상한을 둘 수 있으며, 누군가 만든 스킬은 팀 전체에 공유돼 회사만의 자산으로 쌓인다는 거예요. 누스리서치는 게시물에서 이 공유 스킬이 "회사 고유의 지식재산으로 복리처럼 쌓인다" 고 적었어요. 엔터프라이즈는 같은 기능을 온프레미스나 고객이 고른 클라우드에 올리는 판이고, 회사는 이를 "완전하고 스스로 개선되는 주권형 AI 스택" 이라고 불렀어요.

마케터의 눈으로 보면 이 발표는 제품 설명이 아니라 판매 페이지의 구조를 그대로 옮긴 거예요. 누스 포털의 비즈니스 소개 페이지는 기능을 자동화, 향상, 확산이라는 세 동사로 나눴어요. 첫째는 팀의 누군가가 찾아낸 가장 좋은 스킬이 조직 전체에 자동으로 공유된다는 집단 지혜, 둘째는 모든 AI 모델을 하나의 하네스로 묶어 비용과 투자 대비 효과를 한눈에 본다는 지출 지능, 셋째는 팀이 만든 모든 스킬과 기억과 워크플로가 회사 고유의 역량으로 쌓인다는 자기 영역 소유예요. 페이지는 이 제품을 "팀 구성원 한 사람 한 사람을 위한 스스로 개선되는 에이전트" 라고 소개했어요.

AI

코히어 고메즈, AI 실존 위험 논쟁 공상과학 비판

코히어의 에이단 고메즈 최고경영자가 9월 14일 블룸버그 TV 생방송에서 AI 실존 위험 논쟁이 공상과학 쪽으로 너무 멀리 가 있다고 말했어요. 허깅페이스 해킹 사건을 확대 해석의 예로 들었고, 큰 회사들끼리 규칙을 정하는 건 경고 신호라고 했어요.



1 에이단 고메즈 코히어 최고경영자가 9월 14일 블룸버그 TV 생방송에서 AI 실존 위험 논쟁이 공상과학 쪽으로 너무 멀리 갔다고 말했어요.

2 허깅페이스 해킹 사건은 안전하지 않은 샌드박스에서 시킨 일을 한 것이라 놀랍지 않고, 확대 해석이 너무 멀리 갔다고 주장했어요.

3 안전 기준은 높이되 큰 회사들만 산업 전체의 규칙을 정해선 안 된다며 앞서 낸 카르텔 반박 에세이의 입장을 되풀이했어요.

코히어의 에이단 고메즈 최고경영자가 9월 14일 블룸버그 TV 생방송에서 AI 실존 위험 논쟁이 공상과학 쪽으로 너무 멀리 가 있다고 말했어요. 코히어는 같은 날 이 발언이 담긴 1분 44초짜리 클립을 회사 X 계정에 올렸고, 메탈이 확인한 시점에 조회수는 1만1천 회를 넘겼어요. 진행자가 바로 지난주에 그 이야기가 뉴스였다고 되묻는 자리에서, 프론티어 모델을 만드는 회사의 대표가 카메라 앞에서 논쟁 자체를 비판한 장면이에요.

발언은 길지 않지만 단정적이에요. 고메즈는 "개인적으로 실존 위험 논쟁은 공상과학 쪽으로 너무 멀리 가 있고, 그런 이야기들과 많은 확대 해석에 기대고 있다고 생각합니다" 라고 말했어요. 이어서 "AI가 세상을 장악한다는 생각, 터미네이터 같은 시나리오는 이 단계에서 공론장에 들어와서는 안 된다고 봅니다. 우리는 그 근처에도 가 있지 않고, 앞으로도 그런 일을 내다보지 않습니다" 라고 덧붙였어요. 위험이 작다는 말이 아니라 그 위험을 이야기하는 방식이 틀렸다는 말이에요.

그가 든 사례는 허깅페이스 해킹 사건이에요. 고메즈에 따르면 헤드라인을 뒤덮은 그 사건은 실제로 들여다보면 안전하지 않은 샌드박스 안에 에이전트를 넣고 해킹이라는 과제를 맡긴 것이었고, 에이전트는 시킨 일을 수행해 빠져나왔어요. 그는 "그리 놀랍지 않아요" 라고 했고, 문제는 그 사건에서 끌어낸 확대 해석과 대중에게 전달된 방식이 너무 멀리 갔다는 데 있다고 말했어요. 메탈은 오픈AI의 미공개 모델이 허깅페이스에 침투한 사건을 보도한 바 있으며, 오픈AI가 그 뒤 모니터링과 격리를 강화했다고도 보도한 바 있어요.