AI 뉴스와 용어
월·수·금 오전 7시

METAL
First, AI Magazine

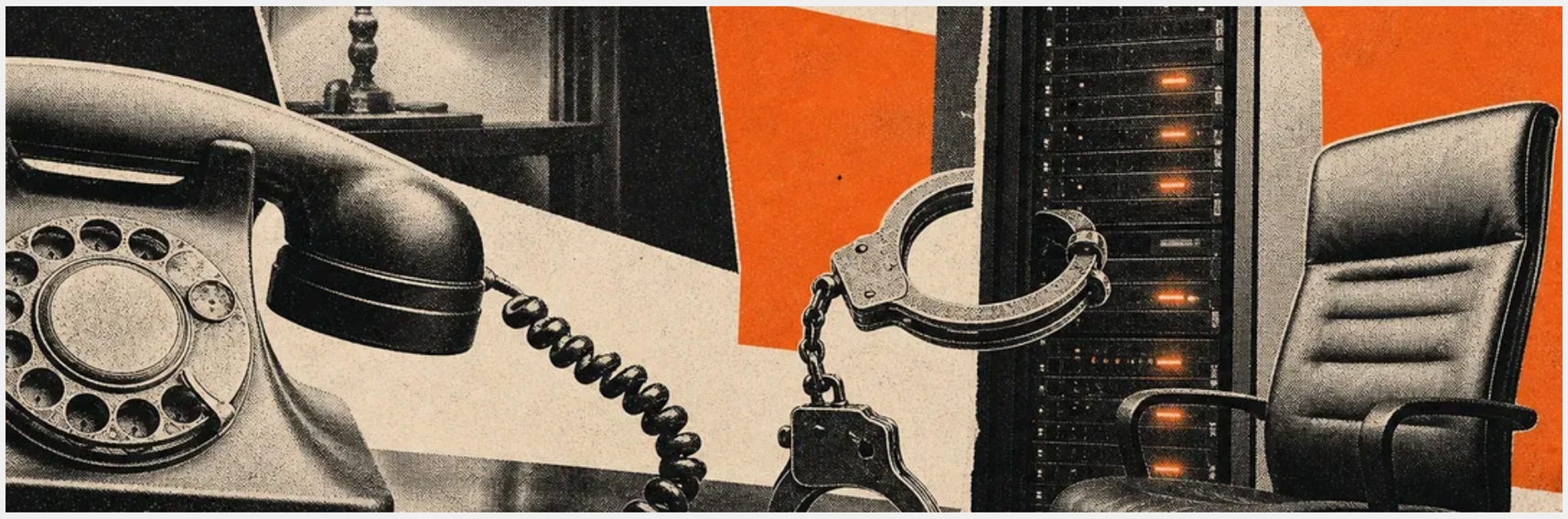
무료 · 언제든 수신거부
metallab.ai

제 62 호 | 2026년 11월 10일 화요일 | 오늘의 AI

AI

트럼프, 대형 AI 기업 수장 투옥 경고 발언

트럼프 대통령이 60 Minutes 인터뷰에서 AI로 나쁜 일을 한다는 보고를 받은 대형 AI 기업 수장에게 법무부가 감옥 갈 준비를 하라고 경고했다고 말했다. 대상과 혐의, 수사 여부는 밝히지 않았고 AI 위험은 법무부가 관리한다고 주장했다.

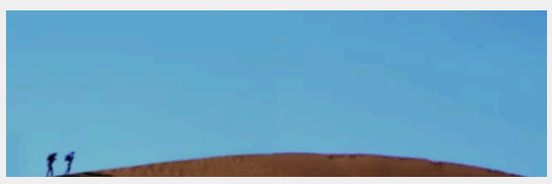


AI · 트럼프, 대형 AI 기업 수장 투옥 경고 발언

도널드 트럼프 미국 대통령이 CBS 시사 프로그램 60 Minutes 인터뷰에서 법무부가 대형 AI 기업 수장에게 투옥 경고를 한 적이 있다고 발언했다. 트럼프 대통령은 AI 위험이 업계가 말하는 만큼 현실적이지 않고 관리할 수 있다며 그 관리 주체로 법무부를 지목했다. 그는 "누군가 AI로 나쁜 일을 하고 있다는 보고가 들어왔고, 우리는 아주 강하게 전화해 감옥 갈 준비를 하라, 오늘 밤 아내에게 잘 자라고 입맞춤해 주라고 했다"고 말했다. 그 인물이 누구인지, 어느 회사인지는 끝내 밝히지 않았다.

인터뷰는 진행자 로스 다우섯이 맡았다. 보도에 따르면 녹화는 금요일 백악관에서 이뤄졌고 편집본이 일요일에 방송됐으며, 중간선거 이후 트럼프 대통령의 첫 대담이었다. 메탈이 확인한 인터뷰 전문에서 다우섯은 중간선거 패배와 이란, 우크라이나 전쟁, 경제를 차례로 물은 뒤 AI로 넘어갔다. 그는 실리콘밸리 사람들이 트럼프 대통령이 AI를 두고 내리는 결정을 가장 오래 기억될 업적으로 꼽는다며, 여론조사에서 미국인은 데이터센터와 AI에 꽤 부정적이라고 짚었다.

투옥 경고 발언은 업계 종사자들의 파국 공포를 왜 걱정하지 않느냐는 질문에서 나왔다. 다우섯은 회사 앞에 서서 인류를 끝장낼 확률이 10%인 회사를 만들고 있다고 말하는 AI 업계 사람들을 무너질지도 모르는 호텔을 자랑하는 건물주에 빗댔다. 트럼프 대통령은 위험은 관리할 수 있다며 "누가 관리하느냐, 법무부다"라고 답했다. 이어 그 인물에게 "지금 당장 집에 가서 아내와 아이들을 보고 잘 자라고 입맞춤해 주라, 내일 아침 감옥에 가니까"라고 말했다고 전했다.



기요금을 낮출 것이라고 주장했다. 다우섯이 백악관에 모였던 AI 거물들 가운데 일론 머스크가 아무도 일하지 않아도 모두가 보편적 고소득을 받는 미래를 말했다고 하자, 트럼프 대통령은 웨스트버지니아처럼 데이터센터를 반기는 지역 사례로 답했다. 발언의 끝에서 그는 "모든 회사를 지켜보는 아주 뛰어난 사람들이 있다"고 강조했다.

메탈은 빅테크 CEO들이 트럼프 대통령의 AI 명칭 변경을 받아들인 흐름과 트럼프 대통령이 AI 포스 창설과 차르 임명을 예고한 소식을 보도한 바 있다. 보도에 따르면 트럼프 대통령은 앞서 인공지능이라는 말을 적들이 쓰는 용어라고 했지만, 이번 인터뷰에서는 SI나 슈퍼 인텔리전스 대신 AI라는 말을 여러 차례 썼다.

기자의 눈으로 보면 이번 발언에서 가장 무거운 대목은 확인되지 않은 것들이다. 미국 대통령이 전국 방송에서 수사기관을 동원해 대형 AI 기업의 수장을 압박했다고 말했는데, 대상과 혐의, 법적 근거, 실제 수사 여부가 하나도 공개되지 않았다. AI 위험 관리를 별도 규제가 아닌 법무부의 재량에 맡긴다는 구상은 규칙이 아니라 개별 압박으로 업계를 다룰 수 있다는 뜻으로 읽힌다. 경고를 받았다는 회사가 어디인지, 무엇을 바꿨는지가 드러나기 전까지 이 발언은 AI 업계 전체가 안고 갈 물음표로 남는다.

AI

누스리서치, 헤르메스 에이전트 보안 플러그인 공개

오픈소스 AI 연구 기업 누스리서치가 11월 8일(미국 시간) 자사 오픈소스 에이전트 헤르메스 에이전트에 코드 저장소의 보안 버그를 찾아 주는 공식 보안 플러그인 hermes-security를 공개했다. 누스리서치는 X에 "헤르메스는 먼저 공격자가 무엇에 닿을 수 있는지, 어떤 입력을 신뢰하는지부터 그린다"며 코드를 병렬 서브에이전트에 나눠 맡기고, 급을 받고, 테스트 러너나 린터처럼 개발할 때만 쓰는 패키지는 낮음으로 센다. 취약한 동작에 실제 코드가 닿는다는 사실이 검토로 드러날 때만 별도 결함으로 올린다.

가장 공을 들인 대목은 증거의 등급이다. 테스트로 결함을 확인하는 선택지를 켜면 에이전트가 결함마다 작은 재현 코드를 쓰고, 저장소의 임시 사본에서 평소 쓰는 도구로 돌린다. 테스트로 확인됨 표시는 그 테스트가 통과하고 짝을 이루는 음성 테스트도 예상대로 동작할 때만 붙는다. 실행되지 못한 테스트는 NOTRUN으로 기록돼 어느 쪽 증거로도 치지 않는다. 테스트는 사용자 권한으로, 자격 증명처럼 보이는 환경 변수를 지운 채 돈다. 문서는 이를 두고 "편의 기능이지 샌드박스가 아니다"라며 믿을 수 있는 코드에서만 테스트를 허용하라고 적었다. 실행 중인 앱을 시험하려면 사용자가 주소를 지정해 허락해야 하고, 그 뒤 한 시간 동안 그 주소에만 최대 200번 요청을 보낸다. 에이전트의 도구는 이 권한을 스스로 줄 수 없다.

보고서도 모델의 말을 그대로 믿지 않게 짜였다. 저장소 안의 텍스트는 에이전트에 대한 지시가 아니라 데이터로만 다루고, 비밀값은 저장하지 않은 채 종류와 위치, 단방향 지문만 남긴다. 깃 훅과 필터는 실행하지 않는다. 결함마다 검사한 스냅샷의 코드를 그대로 인용하고, 끝난 보고서는 봉인돼 나중의 분류 작업이 내용을 바꾸지 못한다. 기각되거나 결론이 나지 않은 후보도 이유와 함께 보고서에 남는다. 수정 패치는 읽기용으로만 보여 주고 플러그인이 적용하지는 않으며, 보고서는 마크다운과 JSON, CSV, SARIF로 내보낸다. 검사가 중간에 멈추면 다음 메시지나 검사 계속 버튼으로 이어 가고, 자동 계속이 진척 없이 세 번 반복되면 일시 정지한다. 데이터는 사용자 프로필 폴더에 저장되지만 검토할 코드는 헤르메스에 설정한 모델 제공사로 건너가므로, 문서는 보고서를 민감한 자료로 다루라고 당부했다.

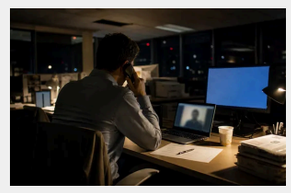
메탈이 확인한 저장소의 평가 문서 EVALUATION.md

오늘 지면

1 AI 트럼프, 대형 AI 기업 수장 투옥 경고 발언
2 AI 쿠켄공업, 70년 도면 검색에 오픈AI 도입
3 AI 메타, 인스타그램 AI 연령 추정 전 세계 확대
4 AI 환경단체·바이커, 구글 데이터센터 반대 청원 전달
5 AI 누스리서치, 헤르메스 에이전트 보안 플러그인 공개
6 AI 앤스로픽 Fable 5.2·Mythos 5.2 출시



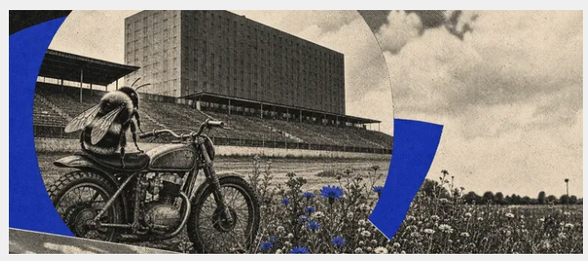
AI 대필 의혹에 집필 기록 도구 등장



미 검찰, AI로 CEO 사칭한 3명 기소

이 지면은 2026년 11월 10일 화요일 아침 구독자 메일함으로 나간 레터를 신문 판형으로 다시 조판한 것입니다. 기사 전문은 metallab.ai 에서 읽을 수 있습니다.

AI



환경단체·바이커, 구글 데이터센터 반대 청원 전달

11월 4일 오후, 영국 곤충 보전 단체 버그라이프의 제이미 로빈스 프로그램 매니저가 런던 킹스크로스에 있는 구글 영국 본사 입구로 걸어 들어갔다. 손에는 런던 동쪽 약 15마일 거리의 서록에 대형 데이터센터를 짓겠다는 계획을 접으라는 청원 내용을 담은 서한이 들려 있었다. 서명자는 5만 5,000명에 가까웠다. 뒤에서는 환경단체와 야생동물 단체 회원들, 그리고 오토바이 경주 클럽 바이커들이 구글에게 다른 곳을 "검색하라"는 손팻말을 들었다.

▶ 2면에 계속

AI

쿠켄공업, 70년 도면 검색에 오픈AI 도입



일본 냉각탑 제조사 쿠켄공업(Kuken Industries)이 70년치 설계 도면을 찾는 검색 도구를 오픈AI API 위에 만들어 도입했다. 오픈AI가 11월 9일 공개한 고객 사례에 따르면 쿠켄공업 엔지니어가 교체 부품의 도면을 찾는 데 걸리는 시간은 1시간 가까이에서 4.5분으로 줄었다. 회사는 1년 기준으로 도면 검색에 드는 시간이 약 90% 줄어들 것으로 내다봤다. 도구는 오픈AI 공식 파트너인 일본 AI 스타트업 리커시브(Recursive)와 함께 만들었다.

쿠켄공업은 대형 건물과 쇼핑센터, 공장에 들어가는 공조·환기 설비를 만든다. 일본 냉각탑 시장에서 가장 큰 점유율을 가진 회사 가운데 하나다. 냉각탑은 순환하는 물의 열을 공기 중으로 내보내 대형 건물과 쇼핑몰은 물론 데이터센터 장비까지 식히는 설비다. 냉각탑에 교체 부품이 필요하면 엔지니어는 70년 동안의 설치 기록을 담은 도면 보관소를 뒤져야 한다. 오늘 만든 부품이 오래된 냉각탑에 맞지 않는 경우가 있어, 원래 설치된 모습을 보여 주는 도면이 없으면 교체 부품을 처음부터 설계하고 제작해야 한다.

AI

메타, 인스타그램 AI 연령 추정 전 세계 확대



메타가 11월 9일(미국 시간) 인스타그램의 AI 연령 추정 기술을 전 세계로 확대한다고 발표했다. 성인 생일을 입력하고 가입한 이용자라도 AI가 10대로 의심하면 10대 계정 보호 설정으로 옮기는 기술이다. 메타는 이번 확대로 전 세계 인스타그램 이용자 가운데 수백만 명의 10대가 더 보호 설정 안으로 들어온다고 밝혔다. 페이스북에는 같은 기술을 계속 배포하고 있고, 스레드에는 연내에 비슷한 기술을 전 세계에 내놓을 계획이다.

10대 계정은 인스타그램과 페이스북, 메신저에 있는 기본 보호 설정이다. 누가 10대에게 연락할 수 있는지, 어떤 콘텐츠를 보는지, 앱에 머무는 시간을 자동으로 제한한다. 보도에 따르면 팔로하거나 이미 연결된 사람에게서만 메시지를 받고, 유해한 댓글을 가리며, 계정이 기본으로 비공개가 된다. 메타는 공지에서 "모든 10대가 이 보호를 받으려면 나이를 알아야 한다"며, 일부 10대가 나이를 속이는 일은 업계 전체의 과제라고 설명했다.

▶ 2면에 계속

◀ 1면에서 계속

AI

쿠켄공업, 70년 도면 검색에 오픈AI 도입



1 일본 냉각탑 제조사 쿠켄공업은 오픈AI 파트너 리커시브와 함께 오픈AI API 기반 도면 검색 도구를 만들어 70년 치 설치 도면을 찾는 데 쓰고 있다.

2 도면 하나를 찾는 시간은 1시간 가까이에서 4.5분으로 줄었고, 회사는 1년 기준 도면 검색 시간이 약 90% 줄어들 것으로 내다봤다.

3 쿠켄공업은 협력사 도면 검토 도구도 개발 중이며, 시제품 시험에서 AI가 짚은 문제의 약 90%가 실제 불일치였다.

문제는 같은 모델명이나 부품명으로도 도면이 여러 장 나온다는 점이었다. 쿠켄공업은 70년 넘게 고객마다 장비를 맞춤 제작해 왔기 때문에 사양이 고객과 설치 시기에 따라 다르다. 보관소가 종이였던 시절에는 검색 한 번에 반나절이 걸리기도 했다. 이후 보관소를 디지털로 옮겼지만 엔지니어는 여전히 어떤 이름이나 번호로 찾아야 하는지, 파일이 어디에 있는지 알고 있어야 했다. 신입 엔지니어는 경험 많은 선배에게 어디를 봐야 하고 도면을 어떻게 읽는지 묻는 일이 잦았다.

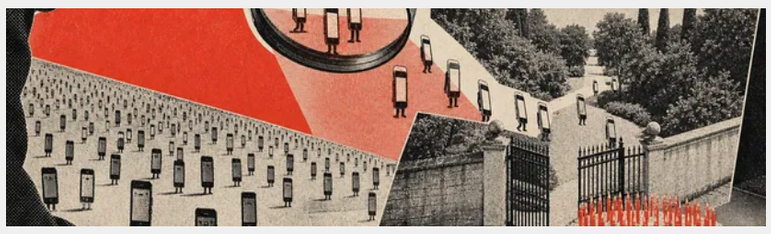
쿠켄공업과 리커시브는 검색 시스템과 현장용 앱을 함께 만들었고, 그 과정에 쿠켄공업의 도면 보관소와 엔지니어들의 의견을 반영했다. 두 회사는 여러 AI 모델에 다양한 도면을 넣어 세부 정보를 얼마나 정확하게 뽑아내는지, 비용은 얼마인지를 시험했다. 오픈AI 모델이 정확도와 비용 두 기준에서 모두 좋은 결과를 냈고, 오픈AI API 덕에 검색 기능을 앱 안에 바로 넣을 수 있었다는 것이 회사 측 설명이다. 메탈이 확인한 사례 영상에서 리커시브 측은 자사 역할을 쿠켄공업의 70년 데이터와 AI 기술을 잇는 일로 소개했고, 오픈AI 모델을 고른 가장 큰 이유로 비정형 데이터에서 정보를 제대로 뽑아낸 내부 벤치마크 결과를 꼽았다.

검색 방식은 현장 엔지니어와의 대화에서 정해졌다. 엔지니어는 도면 번호나 부품 번호를 입력하거나, 도면 또는 부품 목록의 이미지나 PDF를 올리면 된다. 부품 목록 화면을 캡처해 붙여 넣으면 앱이 부품 번호를 읽어 후보 도면을 돌려주기 때문에 번호를 하나씩 손으로 칠 필요가 없다. 결과에는 비슷한 설계의 도면도 함께 나온다. 맞는 도면을 고르는 마지막 판단은 엔지니어가 냉각탑 사양과 과거 교체 기록을 대조해 내린다. 영상에서도 사람이 최종 결정을 내리는 원칙을 지키면서 생각하고 결정하는 시간을 돕는 데 시스템을 쓴다는 설명이 나왔다.

◀ 1면에서 계속

AI

메타, 인스타그램 AI 연령 추정 전 세계 확대



1 메타는 11월 9일 인스타그램의 AI 연령 추정 기술을 전 세계로 확대해 수백만 명의 10대를 10대 계정 보호 설정으로 옮긴다고 발표했다.

2 AI는 생일 축하·학년 언급 같은 맥락 단서와 사진·영상의 시각 단서를 살피며, 메타는 얼굴 인식이나 생체 정보 사용이 아니라고 밝혔다.

3 이번 조치는 초당파 법무장관 52명과의 합의 약속 위에 있고, 메타는 호주에서 애플의 연령대 신호를 쓰기 시작하며 앱스토어 연령 확인 입법을 요구했다.

메탈이 확인한 메타의 5월 5일 공지에 따르면 AI는 프로필 전체에서 생일 축하나 학년 언급 같은 맥락 단서를 찾고, 게시물과 댓글, 소개 글, 캡션을 함께 살핀다. 메타는 같은 공지에서 사진과 영상의 시각 단서를 분석하는 기법을 더하며 "이것은 얼굴 인식이 아니다"라고 강조했다. 보도에 따르면 시각 단서에는 키나 골격 같은 요소가 들어간다. 메타는 7월 31일 이 공지에 정정문을 달아 AI가 생체 정보를 쓰지 않고, 얼굴 지문을 만들거나 사진 속 인물이 누구인지 식별하지 않는다고 밝혔다.

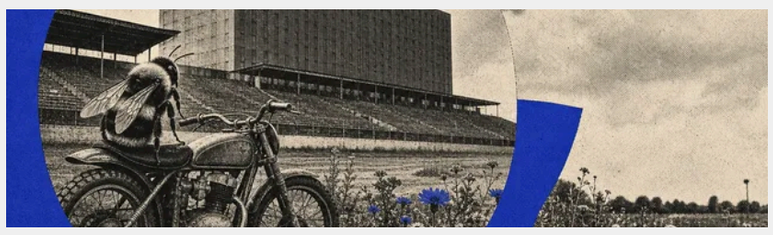
확대는 단계적으로 진행돼 왔다. 메타는 2024년 이후 인스타그램과 페이스북, 메신저에서 수억 명의 10대를 10대 계정에 등록했다. 성인 생일을 쓴 10대를 찾아내는 기술은 지난해 미국과 호주, 캐나다, 영국의 인스타그램에 먼저 적용돼 수백만 개 계정을 보호 설정으로 옮겼다. 5월에는 유럽연합 27개국과 브라질의 인스타그램, 미국 페이스북으로 넓혔고 6월에는 영국과 유럽연합 페이스북으로 이어 갈 계획이었다. 이번 발표로 인스타그램의 적용 범위는 전 세계가 됐다.

13세 미만 이용자에 대한 조치는 더 강하다. 인스타그램과 페이스북은 13세 이상만 쓸 수 있고, AI가 13세 미만으로 판단한 계정은 비활성화된 뒤 나이 증명을 거쳐야 삭제를 면한다. 18세 미만으로 등록된 이용자가 생일을 18세 이상으로 바꾸려 하면 신분증이나 요티의 얼굴 연령 추정 도구로 나이를 확인해야 한다. 메타는 미성년 계정 신고 절차를 단순화하고, 사람 검토팀에 같은 평가 기준을 적용하는 AI 모델을 붙였다고 밝혔다.

◀ 1면에서 계속

AI

환경단체·바이커, 구글 데이터센터 반대 청원 전달



1 버그라이프는 11월 4일 서록 데이터센터 계획 철회를 요구하는 5만 5,000명 가까운 청원 내용을 런던 킹스크로스 구글 영국 본사에 전달했다.

2 곤충·식물 서식지 손실, 연간 최대 57만 톤 이산화탄소 배출, 스피드웨이 경주장 복원이라는 서로 다른 이유로 단체들이 연대했다.

3 구글은 생물다양성 순증과 재생에너지 투자로 상쇄하겠다고 반박했고, 서록 의회는 내년 초 기본 계획 신청을 결정한다.

반대 연대는 이유가 제각각이다. 버그라이프와 지역 야생동물 단체는 희귀 꿀벌과 식물의 서식지가 사라진다고 본다. 비영리단체 폭스글러브는 데이터센터가 해마다 최대 57만 톤의 이산화탄소를 내뿜어 영국 정부의 탄소 예산 안에서 감당할 수 없다고 주장한다. 서록 해머스 오토바이 클럽은 그 자리에서 다시 경주를 하고 싶어 한다. 해머스는 브레이크 없는 500cc 메탄올 오토바이로 흙 트랙을 도는 스피드웨이 팀이다. 보도에 따르면 해머스의 마크 섹스턴 이사는 "우리는 이질적인 존재고 반대하는 근거도 다르다, 그래서 캠페인 전체에 꽤 쓸모가 있다고 생각한다"고 말했다.

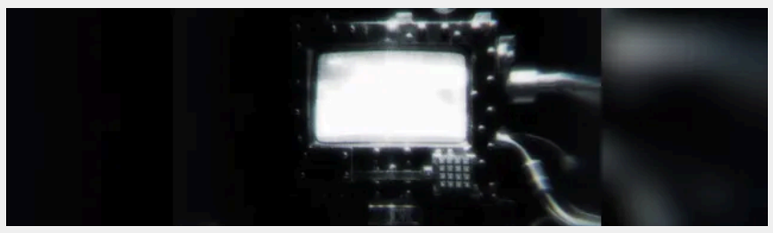
서식지 손실 규모를 두고는 숫자가 맞선다. 버그라이프는 계획대로라면 부지의 핵심 서식지인 옛 개발지 모자이크 서식지의 80% 이상이 사라진다고 분석했다. 메탈이 내려받아 확인한 버그라이프의 2025년 8월 14일 자 2쪽 서한은 개발 신청 서류에 적힌 "보호구역 안 서식지의 약 절반이 사라진다"는 문장을 인용해, 그 넓이가 32에이커가 넘는다고 짚었다. 서한은 구글이 2025년 지속가능성 보고서에서 2024년 말까지 자사 캠퍼스에 약 74에이커의 서식지를 만들거나 복원했다고 자랑한 대목과 이를 나란히 놓았다. 수신인은 케이트 알레시 구글 영국 대표였고, 그때 청원 서명자는 2,500명을 넘긴 상태였다.

주민 여론도 한쪽으로 쏠렸다. 보도에 따르면 계획 당국에 접수된 일반 의견 1,000건 가운데 찬성은 단 1건이었다. 서록 의회는 상당수가 지역 바깥 거주자가 낸 의견이라고 설명했다. 버그라이프와 에식스 야생동물 트러스트, 폭스글러브, 에식스 필드 클럽은 10월부터 지지자들에게 의회 앞 반대 청원을 독려해 왔다. 버그라이프 캠페인 페이지에 따르면 11월 4일 전달 직전 서명은 5만 4,000명을 넘었다.

AI

누스리서치, 헤르메스 에이전트 보안 플러그인 공개

서브에이전트가 코드를 나눠 읽고 공격자가 실제로 닿는 결함만 추려 재현 테스트로 확인하는 공식 보안 플러그인이다. 평가 문서는 정밀도 0.818, 재현율 0.692로 목표치에 못 미친 결과를 그대로 공개했다.



1 누스리서치는 11월 8일 헤르메스 에이전트가 코드 저장소의 보안 버그를 검토하는 공식 보안 플러그인 hermes-security를 공개했다.

2 코드를 약 4,000줄씩 나눠 최대 10명의 서브에이전트가 동시에 읽고, 공격자 도달 가능성을 추적한 뒤 원하면 임시 사본에서 재현 테스트로 결함을 확인한다.

3 합성 사례 26개 평가에서 정밀도 0.818~0.917, 재현율 0.692~0.846으로 목표치에 못 미쳤고, 문서는 출시 품질 주장이 아니라고 밝혔다.

오픈소스 AI 연구 기업 누스리서치가 11월 8일(미국 시간) 자사 오픈소스 에이전트 헤르메스 에이전트에 코드 저장소의 보안 버그를 찾아 주는 공식 보안 플러그인 hermes-security를 공개했다. 누스리서치는 X에 "헤르메스는 먼저 공격자가 무엇에 닿을 수 있는지, 어떤 입력을 신뢰하는지부터 그린다"며 코드를 병렬 서브에이전트에 나눠 맡기고, 후보 결함마다 실제 코드와 대조한 뒤 공격자가 정말 악용할 수 있는지 추적한다고 밝혔다. 플러그인은 X 계정 @marks가 만들었고 채팅과 명령줄, 헤르메스 데스크톱의 보안 페이지 세 곳에서 돌릴 수 있다. 결과 보고서는 보안 도구들이 주고받는 표준 형식인 SARIF로도 내보낸다.

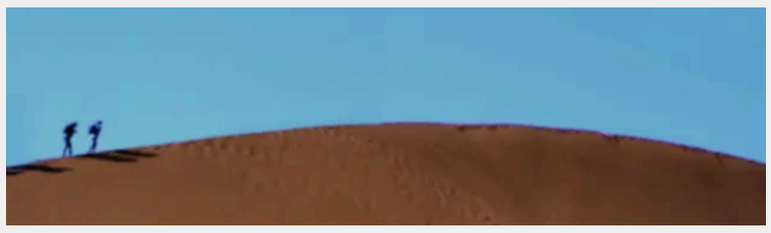
기존 정적 분석 도구는 위험해 보이는 코드 패턴을 찾아내지만, 그 코드에 공격자가 실제로 닿는지는 사람이 따로 따져야 했다. hermes-security는 이 도달 가능성 질문을 검토의 한가운데 두고, 별도 스캐너 없이 모델이 직접 코드를 읽는다. 설치에는 헤르메스 0.21.5 이상과 파이썬 3.11 이상이 필요하고 명령 한 줄이면 켜진다. 데스크톱 보안 페이지는 다른 데스크톱 플러그인처럼 꺼진 채로 시작해 사용자가 직접 켜야 한다. 보도에 따르면 누스의 공식 플러그인 가운데 정적 분석 작업을 겨냥한 것은 이번이 처음이다. 메탈은 누스리서치가 헤르메스 플러그인 카탈로그를 연 소식과 헤르메스 에이전트 v0.23.0 공개를 보도한 바 있으며, 이번 플러그인도 누스리서치의 공식 플러그인 묶음에 들어 있어 명령 한 줄로 설치된다.

검사는 저장소를 폴더 단위로 약 4,000줄씩 자르는 데서 시작한다. 리뷰어 한 명이 한 번에 읽을 수 있는 분량이라는 설명이다. 리뷰어는 최대 10명이 동시에 돌고, 한 명이 더 붙어 전체 범위에서 진입점과 신뢰 경계를 살핀다. 표준 검사는 소스와 설정, CI, 인프라 코드, 마이그레이션 같은 운영 코드만 읽고 테스트와 문서, 생성 파일은 건너뛰되 그 안에 커밋된 비밀값이 있으면 예외로 읽는다. 운영 코드가 90청크, 약 36만 줄 안에 들어오면 전부 읽고, 더 크면 라우트와 요청 입력, 셸 명령, SQL, 역직렬화, 인증, 웹훅을 다루는 파일부터 순위를 매겨 예산이 다할 때까지 읽는다. 가장 위험한 청크 최대 2개에는 두 번째 독립 리뷰어가 붙고, 끝내지 못한 파일은 새 리뷰어에게 넘어가 최대 여섯 라운드까지 이어진다. 심층 검사는 건너뛰기 목록 없이 저장소 전체를 기본 세 번 독립적으로 훑고, 커밋하지 않은 작업이나 브랜치 변경분만 보는 선택지도 있다. 8MB가 넘는 파일과 벤더 코드, 심볼릭 링크는 건너뛰고 보고서에 그 목록을 남긴다.

AI

앤스로픽 Fable 5.2·Mythos 5.2 출시

같은 기반 모델에 안전장치를 달리한 두 플래그십으로, 출력 속도는 최대 30% 빨라지고 가격은 내려갔다. 에이전트 1,000개 팀이 페르마의 마지막 정리 증명을 30시간 만에 형식화했다.



1 앤스로픽은 11월 9일 Claude Fable 5.2와 Claude Mythos 5.2를 출시했고, Fable은 모든 사용자에게, Mythos는 검증 프로그램을 통과한 기관에만 제공한다.

2 Fable 5.2는 FrontierCode 2.0에서 49.1%로 Opus 5.5의 45.5%를 넘었고, RoboDojo Sim에서는 64.9%로 GPT-6 Astra의 두 배를 넘었다.

3 Fable 5.2 에이전트 1,000개 팀은 페르마의 마지막 정리를 30시간 만에 형식화했고, Mythos 5.2는 CVE Rediscovery에서 Opus 5.5보다 10.2% 많은 취약점을 찾았다.

앤스로픽이 11월 9일(미국 시간) 플래그십 모델 Claude Fable 5.2와 Claude Mythos 5.2를 출시했다. 두 모델은 같은 기반 모델이다. Fable은 사이버보안과 생물학 오남용을 막는 강한 안전장치를 달아 모든 사용자에게 열었고, Mythos는 생명과학·사이버 검증 프로그램을 통과한 기관에만 제공한다. 앤스로픽은 출시 전 시험에서 두 모델이 엔지니어들이 엄두를 못 내던 소프트웨어 이전 작업을 끝내고, 전자기기 세대를 처음부터 설계하고, 블랙홀 물리학의 미해결 문제 두 개를 풀고, 페르마의 마지막 정리 증명을 30시간 만에 형식화했다고 밝혔다. 출력 속도는 5.1보다 최대 30% 빨라졌고, 가격은 입력 100만 토큰당 8달러, 출력 100만 토큰당 40달러로 낮아졌다.

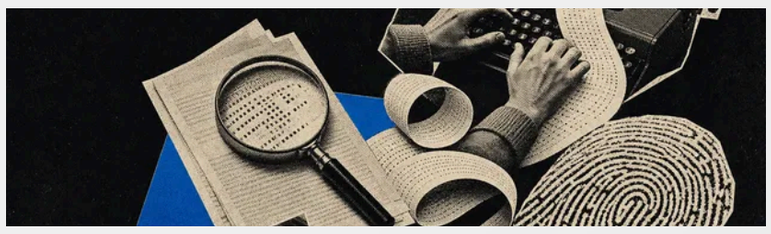
선택 사항인 고속 모드는 토큰 생성 속도를 최대 2.5배 끌어올리며 입력 100만 토큰당 16달러, 출력 80달러다. 앤스로픽은 Fable과 Mythos를 가장 야심 차고 열린 기술 작업용으로 규정했다. 대부분의 코딩과 지식 작업은 일상 기본 모델인 Claude Opus 5.5로 시작하고, 작업 범위와 복잡도에 따라 Fable이나 Mythos로 올리거나 Sonnet 5.5, Haiku 5.5로 내려가라는 것이 회사의 권고다. Fable 5.2는 작은 모델이 어려운 문제를 물어보는 어드바이저, 쉬운 일을 작은 모델에 나눠 맡기는 오케스트레이터 역할도 맡는다. 시험 참여자들은 Fable 5.2와 Mythos 5.2가 이전 세대보다 더 명확하게 소통한다고 평가했다.

메탈이 확인한 앤스로픽의 벤치마크 표에서 Fable 5.2는 평가한 거의 모든 항목에서 Opus 5.5를 앞섰다. 오픈소스 유지관리자가 모델의 코드 변경을 실제로 병합할지를 재는 FrontierCode 2.0에서 Xhigh 노력 기준 49.1%를 기록해 Fable 5.1 32.4%, Opus 5.5 45.5%, GPT-6 Astra 42.8%를 넘었다. High 노력에서는 45.3%로 Opus 5.5 최고점과 비슷한 점수를 과제당 약 절반 비용으로 냈다. FrontierSWE v2는 69.9%, 도구를 쓰는 인류의 마지막 시험은 74.2%였고, 원격 노동 지수는 41.3%로 Fable 5.1의 17.9%에서 두 배 넘게 올랐다. 업무 자동화 벤치마크 AutomationBench에서는 51.4%로 Gemini 4 Carbon의 53.9%에 못 미쳤고, 지식 노동 평가 AA-Briefcase의 Elo 점수도 1785로 Opus 5.5의 1800보다 낮았다.

AI

AI 대필 의혹에 집필 기록 도구 등장

AI가 썼다는 의혹에 맞서 집필 과정을 기록하고 원고에 암호 지문을 새기는 도구가 잇따라 나왔다. 작가조합과 탐지 회사, 출판 전문가는 이 도구가 의혹의 해법이 되기 어렵다고 짚었다.



1 워킹스·블라인드프루프·윈터텍스트·휴먼잉크·오브라베라 같은 집필 기록 도구가 작가의 집필 과정을 추적하고 원고에 암호 지문을 새겨 저작 기록을 남긴다.

2 3월 공포 소설 「샤이 걸」이 AI 대필 의혹 끝에 출간 취소됐고, 판그램은 본문 약 78%를 AI 생성으로 판정했지만 작가는 의혹을 부인했다.

3 작가조합 메리 라센버거 최고경영자는 기록을 챙기라는 요구가 모욕적이라고 했고, 판그램과 오브라베라 측도 도구가 AI 사용 여부까지 증명하지는 못한다고 밝혔다.