AI 뉴스와 용어
월·수·금 오전 7시

METAL
First, AI Magazine

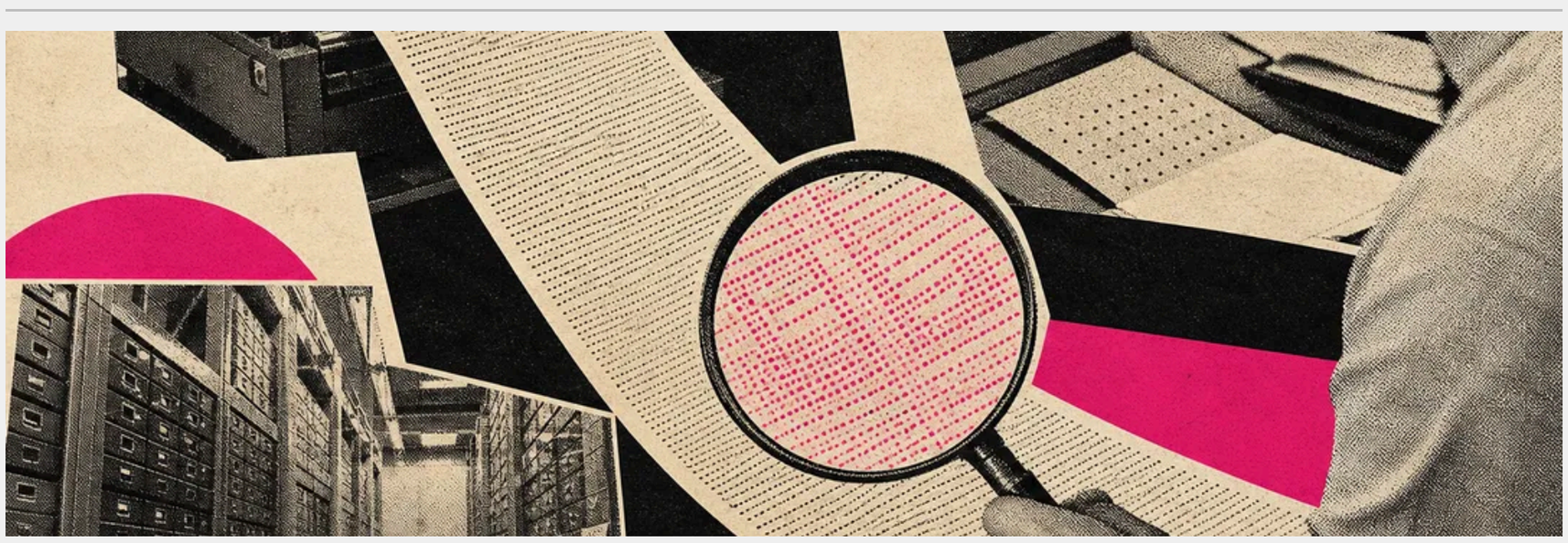
무료 · 언제든 수신거부
metallab.ai

제 38 호 | 2026년 10월 6일 화요일 | 오늘의 AI

AI

오픈AI, EU 텍스트 워터마크 도입 발표

오픈AI가 EU 인공지능법에 맞춰 챗GPT와 코덱스 출력에 보이지 않는 텍스트 워터마크 textGrain을 넣고 API 고객에게는 선택 기능으로 연다. 낱말 25%만 바꿔도 판별률이 17%로 떨어진다는 수치를 함께 공개하고 판별기는 연구자에게만 열었다.



AI · 오픈AI, EU 텍스트 워터마크 도입 발표

오픈AI가 10월 5일(현지시간) 유럽연합(EU) 인공지능법(AI Act)의 텍스트 출처 규칙에 대응하는 텍스트 워터마크 도입 계획을 발표했다. 회사가 직접 만든 워터마크 기술의 이름은 textGrain이다. 유럽연합 안에서는 앞으로 몇 주에 걸쳐 챗GPT와 코덱스(Codex)가 쓰는 글 가운데 적용 대상 출력에 눈에 보이지 않는 워터마크가 붙고, 전 세계 API 고객은 발표 당일부터 일부 모델에서 워터마크를 직접 켤 수 있게 됐다. API의 기본값은 꺼짐이다. 워터마크를 찾아내는 판별기는 일반에 풀지 않고, 심사를 통과한 연구자와 전문 기관에만 신청을 받아 열기로 했다.

EU 인공지능법은 생성형 AI 제공자에게 AI가 만든 텍스트를 기계가 읽을 수 있는 방식으로 식별 가능하게 하라고 요구한다. 보도에 따르면 이 의무를 담은 제50조는 2026년 8월 2일 발효됐고, 기존 시스템에는 12월 2일까지 전환 기간이 주어졌다. 오픈AI는 발표문에서 "텍스트 워터마킹과 판별은 여전히 초기 기술이며 상당한 한계를 안고 있다"고 밝히고, 단계적 접근이 법의 요구와 기술의 현실을 함께 반영한 결과라고 설명했다. 오픈AI는 앞서 자사 지원 페이지에 "텍스트를 포함한 모든 형식으로 출처 신호를 넓히는 것이 목표"라고 적은 바 있다.

textGrain은 모델이 다음 낱말을 고르는 확률에 비밀 키로 만든 통계적 신호를 섞는 방식이다. 읽는 사람 눈에는 아무 차이가 없지만, 같은 키를 가진 판별기는 낱말 선택에 남은 치우침을 찾아 오픈AI 워터마크가 들어 있는지 가린다. 메탈이 확인한 20쪽 분량의 기술 보고서에 따르면 이 방법은 어휘를 키에 따라 여러 묶음으로 나누고, 최적 수송(optimal transport) 문제를 풀어 어떤 묶음이 뽑힐지를 키와 엮는다. 묶음 안에서는 낱말끼리의 원래 확률 비율을 그대로 둔다. 키를 무작위로 뽑아 평균을 내면 원래 모델의 분포가 그대로 돌아오도록 설계해, 워터마크가 모델의 답을 한쪽으로 몰지 않게 했다.

엔지니어 눈으로 볼 때 핵심 장치는 엔트로피 예산이다. 기존 굼벨-맥스(Gumbel-max) 방식은 같은 문맥과 같은 키에서 늘 같은 낱말을 고르기 때문에, 같은 질문을 여러 번 던지면 똑같은 답이 반복될 수 있다. 보고서는 이 문제를 짚으며 워터마크가 앗아 가는 샘플링 무작위성을 KL 발산 값 하나로 재고, 그 상한을 엔트로피의 일정 비율로 묶었다. 판별기는 생성된 텍스트와 비밀 키만 있으면 돌아가고, 생성에 쓴 모델이나 예산 값을 알 필요가 없다. 보고서는 이 방식이 초안 모델과 본 모델이 키를 공유하는 투기적 샘플링(speculative sampling)과도 함께 쓸 수 있다고 밝혔다. 저자는 9명으로 오픈AI 연구진 5명과 펜실베이니아대·예일대 연구진 4명이며, 교신저자는 오픈AI의 웨이제 수(Weijie Su)다.

오픈AI가 공개한 성능 수치는 조건에 따라 크게 갈렸다. 오탐률 목표를 1%로 둔 시험에서 판별기는 심리학 분야 답변 200토큰 분량의 약 80%, 400토큰 분량의 약 95%에서 워터마크를 찾았다. 낱말 선택의 폭이 좁은 수학 답변에서는 판별률이 크게 떨어져, 회사가 공개한 그래프상 200토큰에서 40%에 못 미쳤고 400토큰에서도 60% 안팎에 머물렀다. 손을 댄 글은 더 어려웠다. 400토큰 글에서 낱말의 10%를 동의어로 바꾸자 판별률은 약 92%에서 66%로 내려갔고, 25%를 바꾸자 17%까지 떨어졌다. 오픈AI는 자사가 시험한 다른 방식들, 구글의 텍스트용 SynthID와 비교해도 textGrain이 같거나 나은 성능을 냈다고 밝혔다. 다만 이상적 조건의 성적이 일상 사용에서의 안정적 판별을 보장하지 않는다는 점도 발표문에 함께 적었다.

품질 손실은 거의 없었다는 것이 회사 설명이다. 오픈AI는 최신 프런티어 모델 아스트라(Astra)를 최대 설정으로 돌려 워터마크를 넣은 출력과 넣지 않은 출력을 8개 벤치마크에서 비교했다. 아티피셜 애널리시스 인텔리전스 지수는 49.57점에서 49.76점, GPQA 다이아몬드는 94.44%에서 93.94%, DeepSWE v11은 72.80%에서 71.68%. 터

따라 사례별로 접근을 허용하며, 오픈AI 워터마크가 있는지만 알려 주고 이용자가 누구인지나 프롬프트와 대화 내용은 드러내지 않는다. 회사는 워터마크 결과로 알 수 없는 것도 다섯 가지로 못 박았다. 사람이 얼마나 기여했는지, 글의 소유권과 책임이 누구에게 있는지, 누가 썼는지, 내용이 정확한지를 워터마크는 말해 주지 않는다. 워터마크가 안 잡혔다고 사람이 쓴 글이라는 증거도 아니다. 너무 짧거나 고쳐 쓰였거나 번역됐을 수 있고, 다른 회사 도구로 만든 글일 수도 있기 때문이다.

경쟁사 중에서는 앤스로픽이 먼저 움직였다. 메탈은 앤스로픽이 클로드가 쓴 글에 보이지 않는 워터마크를 넣기로 한 소식을 보도한 바 있다. 앤스로픽은 당시 "첫날부터 AI 생성 콘텐츠에 표시하겠다"고 밝히며 지원 지역 전체에 적용했다. 오픈AI는 반대로 챗GPT 워터마크를 유럽연합에서만 켜고 세계 기본값으로 삼지 않는다고 밝혔다. 지역 단위로 시작해 실제 사용과 피드백에서 배우겠다는 이유다. 보도에 따르면 마이크로소프트, 메타, 미스트랄도 같은 강령에 서명했지만 xAI는 서명하지 않아 그록(Grok)에는 같은 약속이 없다.

이미지와 오디오 쪽 장치는 그대로 유지된다. 오픈AI는 지원되는 이미지에 콘텐츠 자격 증명(Content Credentials)을 붙이고 C2PA 표준을 따르며, 이미지와 오디오에 구글의 SynthID 워터마크를 넣는다. openai.com/verify 웹 도구와 콘텐츠 출처 API(Content Provenance API)도 계속 공개로 열어 둔다. 회사는 앞으로 몇 주 안에 클라우드 파트너를 거쳐 쓰는 오픈AI 모델 출력에도 워터마크를 제공하고, 기술 보고서를 보강하며, textGrain을 오픈소스로 내놓을 계획이라고 밝혔다.

이번 발표가 남긴 숫자는 25%다. 낱말 넷 중 하나를 바꾸면 판별률이 17%로 떨어진다는 사실은, 텍스트 워터마크가 AI 글을 잡아내는 탐지기가 아니라 손대지 않은 출력

오늘 지면

1 AI 오픈AI, EU 텍스트 워터마크 도입 발표
2 AI 오픈AI, 챗GPT 시각 광고 형식 시험
3 AI 웨이페어, 오픈AI로 사양 1,100만 건 검증
4 비즈니스 샘 올트먼, 앤스로픽과 규제 세계관 차이 강조
5 AI 코히어, 기업용 에이전트 플랫폼 North 2 공개

이 지면은 2026년 10월 6일 화요일 아침 구독자 메일함으로 나간 레터를 신문 판형으로 다시 조판한 것입니다. 기사 전문은 metallab.ai 에서 읽을 수 있습니다.

비즈니스



샘 올트먼, 앤스로픽과 규제 세계관 차이 강조

샘 올트먼 오픈AI 최고경영자가 AI 규제를 두고 앤스로픽과 근본적인 세계관 차이가 남아 있다고 강조했다. 올트먼은 미국의 한 정치 전문 매체가 10월 5일(현지시간) 창간한 기술 뉴스레터 ·팟캐스트의 첫 호 단독 인터뷰에서 AI의 혜택을 위해 일부 위험은 받아들여야 하고 기술은 대중이 폭넓게 쓸 수 있어야 한다고 주장했다. 두 회사의 정책 거리가 최근 몇 달 사이 좁혀진 가운데 나온 발언이라, 올트먼이 수렴 속에서도 선을 긋는 데 공을 들였다는

▶ 2면에 계속

AI

오픈AI, 챗GPT 시각 광고 형식 시험



오픈AI가 챗GPT 안에 이미지를 앞세운 새 시각 광고 형식을 시험한다. 오픈AI는 10월 5일(현지시간) 공식 블로그에서 이 광고를 챗GPT의 이미지 생성 과정에 먼저 붙이고, 이달 말 미국에서 초기 광고주 그룹과 시험을 시작한다고 밝혔다. 같은 발표에서 오픈AI는 광고 성과 측정 도구와 측정 파트너를 넓히고, 외부 기관과 함께 브랜드 적합성 평가 시범 사업을 준비한다고 공개했다. 광고 지면을 하나 더 여는 동시에 광고주가 예산을 옮기기 전에 묻는 질문, 곧 효과를 어떻게 증명하고 어떤 대화 옆에 광고가 붙는지에 한꺼번에 답하려는 발표다.

오픈AI는 챗GPT가 매주 12억 명에게 닿는다며 스스로를 세계 최대의 AI 네이티브 소비자 플랫폼이라고 규정했다. 회사는 광고가 구독만으로는 닿지 못하는 더 큰 시장을 상대하게 해 주는 사업의 강점이자, AI의 혜택을 더 많은 사람에게 넓히는 수단이라고 설명했다. 메탈은 오픈AI가 챗GPT 광고에 스폰서 에이전트를 시험한 소식과 챗GPT 광고를 동남아시아와 대만으로 확대한 소식을 보도한 바 있다.

AI

웨이페어, 오픈AI로 사양 1,100만 건 검증



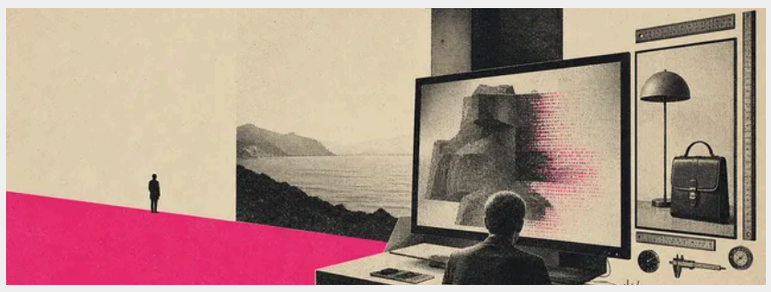
미국 가구·생활용품 전자상거래 기업 웨이페어가 오픈AI 모델로 제품 사양 1,100만 건 이상을 검증하고 매달 공급업체 지원 티켓 4만 1,000건을 자동으로 처리하고 있다고 오픈AI가 9월 25일 공개한 고객 사례에서 밝혔다. 일부 업무 흐름에서는 전체 티켓의 최대 70%가 사람 손을 거치지 않고 처리된다. 웨이페어는 약 1만 2,000명 직원에게 챗GPT 엔터프라이즈 1,200석도 배포했다. 새 상품을 늘리는 대신 이미 가진 약 4,000만 개 상품의 정보를 바로잡아 매출 기회를 키우겠다는 것이 이 사례의 골자다.

출발점은 상품 정보의 품질이었다. 웨이페어 카탈로그 팀은 1,000개 가까운 제품 분류에 걸쳐 수천만 개 상품을 관리하고, 색상·소재·크기 같은 태그가 검색과 추천, 상품 진열을 움직인다. 지원해야 하는 태그만 4만 7,000종이다. 오픈AI와 손잡기 전에는 공급업체나 고객이 무언가 잘못됐다고 알려와야 고칠 수 있었고, 사람의 손으로는 물량을 따라잡지 못했다. 회사가 먼저 만든 맞춤형 태깅 AI 모델은 효과가 있었지만 규모가 커질수록 비용이 감당하기 어려웠다.

◀ 1면에서 계속

AI

오픈AI, 챗GPT 시각 광고 형식 시험



1 오픈AI가 10월 5일 챗GPT 이미지 생성 과정에 붙는 새 시각 광고 형식을 이달 말 미국에서 초기 광고주와 시험한다고 밝혔다.

2 하이터치·틸리움·라이브램프 연동과 앱스플라이어 등 10개 기여도 분석 파트너를 더해 광고 성과 측정 도구를 넓혔다.

3 더블베리파이·IAS와 실제 대화에 접근하지 않는 브랜드 적합성 평가 시범 사업을 개발하고, 네거티브 프레이즈 기능을 열었다.

새 시각 광고 형식은 제품에서 얻는 영감, 제품을 쓰는 장면, 그 제품이 가능하게 하는 경험을 이미지로 보여 준다. 오픈AI는 이용자가 새 브랜드를 발견하고, 무엇이 그 상품을 다르게 만드는지 이해하고, 무엇을 살지 정하는 데 이 형식이 쓰인다고 밝혔다. 광고는 분명하게 광고로 표시되고, 이용자가 만들고 있는 이미지와는 분리된다. 오픈AI는 다른 광고 형식과 마찬가지로 광고가 챗GPT의 답변에 영향을 주지 않는다고 덧붙였다.

측정 쪽 변화가 더 크다. 오픈AI는 자체 측정과 파트너 주도 측정을 함께 지원한다는 원칙 아래 하이터치, 틸리움, 라이브램프와의 연동을 붙여 광고주가 기존 시스템의 전환 데이터를 챗GPT 광고로 보내기 쉽게 했다. 전환 API와 리포팅, 클릭 기여도 분석을 지원하는 파트너로는 앱스플라이어, 트리플 웨일, 어저스트, DV 로커박스, 노스빔, 브랜치, 싱귤러, 코차바, 에어브릿지, 텐진 등 10곳을 꼽았다. 풀퍼널·고급 측정 파트너로는 포스파, 메저드, 인크리멘털이 이름을 올렸다.

광고가 없었어도 일어났을 구매를 걸러 내는 증분 측정은 아직 초기 단계라고 오픈AI는 스스로 밝혔다. 회사는 하우스, 메저드, 워크매직과 손잡고 지역 단위로 광고 노출을 나눠 비교하는 실험을 탐색해 챗GPT 광고의 인과 효과를 보여 주겠다는 계획이다.

◀ 1면에서 계속

AI

웨이페어, 오픈AI로 사양 1,100만 건 검증



1 오픈AI가 9월 25일 공개한 고객 사례에 따르면 웨이페어는 오픈AI 모델로 상품 사양 1,100만 건 이상을 검증·수정했다.

2 자체 공급업체 지원 플랫폼 윌마는 매달 티켓 4만 1,000건을 자동 처리하며 일부 업무 흐름에서는 자동화 비율이 최대 70%에 이른다.

3 카탈로그 품질 시스템을 적용한 A/B 테스트에서 노출·클릭·페이지 순위가 유의미하게 올랐고, 웨이페어는 직원 약 1만 2,000명에게 챗GPT 엔터프라이즈 1,200석을 배포했다.

제시카 다시 웨이페어 카탈로그 머천다이징 어소시에이트 디렉터는 "데이터 품질이 좋아질수록 고객과 쌓는 신뢰도 커진다"며 "잘못 표시된 상품 때문에 생기는 반품 같은 비싼 후속 문제를 직접 줄여 준다"고 말했다. 상품 설명이 틀리면 고객은 기대와 다른 물건을 받고, 그 비용은 반품으로 돌아온다.

웨이페어가 오픈AI 모델로 만든 카탈로그 품질 시스템은 두 단계로 일한다. 먼저 웹 정보와 웨이페어 내부 정의를 바탕으로 각 라벨이 무엇을 뜻하는지 설명하는 지침서를 만든다. 그다음 모델이 그 지침서를 기준으로 상품 정보를 읽고 알맞은 라벨을 붙인다. 이 방식 덕에 웨이페어는 AI가 점검할 수 있는 상품 속성의 범위를 지난해보다 70배 빠른 속도로 넓히고 있다고 회사 측은 설명했다.

메탈이 확인한 오픈AI 공개 화면은 이 시스템이 실제로 무엇을 고치는지 보여 준다. 「원목 원형 커피 테이블」이라는 상품의 제목에는 월넛이 적혀 있지만 상품 설명에는 서던 옐로 파인이라고 돼 있어, AI는 수종을 소나무로 바로잡고 월넛은 수종이 아니라 착색 색상이라는 근거를 달았다. 다리 모양과 마감 여부, 가장자리 형태도 사진과 대조해 고쳤고, 비어 있던 서랍 항목은 새로 채웠으며, 치수와 상판 두께는 도면과 맞는지 확인만 했다. 한 상품에서 수정 4건, 항목 추가 1건, 검증 2건이 나왔다.

◀ 1면에서 계속

비즈니스

샘 올트먼, 앤스로픽과 규제 세계관 차이 강조



1 샘 올트먼 오픈AI 최고경영자가 10월 5일 공개된 단독 인터뷰에서 앤스로픽과 AI 규제를 둘러싼 근본적 세계관 차이가 남아 있다고 강조했다.

2 올트먼은 기술의 혜택을 위해 일부 나쁜 일은 받아들여야 한다면서도 심각한 통제 상실 같은 파국적 위험은 받아들이지 않는다고 밝혔다.

3 앤스로픽은 자사 규제 제안이 프런티어 모델에만 적용된다고 반박했고, 두 회사는 속도 조절과 외부 평가자 상주 등에서는 이미 같은 쪽에 서 있다.

올트먼은 반대편 논리도 직접 옮겼다. 그는 "동의하지는 않지만, 이 기술이 너무 강력하고 위험해져서 샌프란시스코의 연구소 하나가 그것을 쥐고 나쁜 일이 일어나지 않게 하면서 혜택을 나눠 줘야 한다고 생각하는 사람들의 관점은 이해한다"고 말했다. 그러고는 그런 방식을 "전혀 받아들일 수 없는 거래"라고 부르며 오픈AI가 지지하는 가벼운 규제 기조와 맞지 않는다고 했다. 보도에 따르면 이 발언은 도널드 트럼프 대통령과 가까운 기술 자문가들이 앤스로픽을 겨냥해 써 온 규제 포획 비판을 끌어온 것이다.

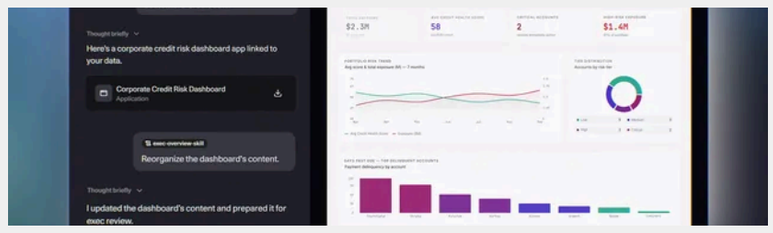
다만 올트먼이 모든 위험을 감수하자고 한 것은 아니다. 그는 이후 AI에 대한 심각한 통제 상실을 포함한 정말 파국적인 위험은 받아들이지 않는다고 덧붙였다. 보도에 따르면 올트먼은 앞선 다른 인터뷰에서도 이번 10년 안에 AI가 인류를 멸망시킬 위험이 10%라면 받아들일 수 없다고 했고, "어떤 연구소도 정렬 문제를 풀지 못했다고 믿는다"고 말한 바 있다.

정책의 실제 거리는 좁혀져 왔다. 보도에 따르면 앤스로픽은 오랫동안 업계에서 가장 강한 AI 규제 지지자로 여겨졌고, 오픈AI는 최근까지 더 가벼운 접근을 선호했다. 그러나 오픈AI는 자사 모델이 얽힌 해킹 사건과 모델의 일탈 행동에 대응하면서 더 강한 규칙을 지지하는 쪽으로 움직였다. 올트먼은 지난달 가장 앞선 AI 모델의 개발 속도를 늦추자는 아모데이의 제안에 동의했고, 오픈AI는 앤스로픽을 따라 이전보다 엄격한 안전 요건을 담은 새 주법들을 지지했다. 오픈AI 로비스트들은 최고 AI 기업에 외부 안전 평가자를 상주시키도록 하는 미 하원의 초당적 법안에도 지지를 표했다. 메탈은 아모데이가 AI 개발 속도 조절과 외부 평가자 상주를 제안한 에세이와 올트먼이 반독점 면제 없이 안전 조치를 하겠다고 밝힌 성명을 보도한 바 있다.

AI

코히어, 기업용 에이전트 플랫폼 North 2 공개

코히어가 10월 5일 오케스트레이션을 새로 설계하고 메모리와 토큰 지출 상한을 넣은 기업용 AI 에이전트 플랫폼 North 2를 공개했다. 앤스로픽·오픈AI·구글 등 다른 회사 모델도 꽂아 쓸 수 있게 플랫폼을 열었다.



1 코히어가 10월 5일 15개가 넘는 새 기능을 담은 기업용 AI 에이전트 플랫폼 North 2를 공개했다.

2 새 오케스트레이션 시스템과 세션 간 메모리, 스킬·라이브러리를 넣었고, 관리 콘솔 North Admin에서 사용자·에이전트 단위로 토큰 지출을 추적하고 상한을 건다.

3 코히어 모델 외에 다른 회사 모델도 쓸 수 있으며, LG CNS·벨 사이버·래디컬 벤처스가 고객 사례로 소개됐다.

캐나다 AI 기업 코히어가 10월 5일 기업용 AI 에이전트 플랫폼 North 2를 공개했다. 여러 단계로 이어지는 에이전트 작업을 조율하는 오케스트레이션 시스템을 새로 설계했고, 세션이 바뀌어도 맥락을 기억하는 메모리와 관리자가 토큰 지출에 상한을 거는 비용 통제 기능을 넣었다. 코히어는 공식 X 계정에서 이번 판을 15개가 넘는 새 기능을 담은 「가장 큰 업그레이드」라고 소개했다. 코히어 모델만이 아니라 고객이 가져온 다른 회사 모델도 돌릴 수 있게 플랫폼을 열었다는 점이 가장 큰 변화다.

North는 코히어가 2025년 1월 대표 고객인 캐나다 은행 RBC와 함께 처음 공개하고 그해 8월 정식 출시한 플랫폼이다. 코히어는 지난 1년 동안 금융, 의료, 통신, 제조, 에너지, 공공 부문에 North를 배치하며 대규모 실제 운영 경험을 쌓았고, 그 경험이 North 2의 설계를 정했다고 밝혔다. 회사는 기업들이 에이전트를 도입하면서 보안이 충분하면 똑똑하지 않고, 똑똑하면 통제가 안 되고, 통제가 되면 운영비가 너무 비싸지는 선택을 강요받아 왔다며 North 2가 그 맞바꿈을 끝낸다고 주장했다.

새 오케스트레이션 시스템은 여러 단계 작업을 스스로 처리하다가 사람의 승인이 필요한 단계에서만 사람을 부른다. 간단한 프롬프트로 만든 에이전트와 자동화는 회사 전체에 공유돼, 다른 팀이 같은 업무 흐름을 처음부터 다시 만들 필요가 없다. 에이전트가 반복해 꺼내 쓰는 로직은 스킬로, 모든 에이전트가 참고할 문서와 자산은 공유 라이브러리로 묶는다. 대화창 안에서 자연어로 슬라이드, 대시보드, 문서, 간단한 앱의 시제품을 만들 수 있고, 반복 업무에는 미리 만든 템플릿과 끌어다 놓는 방식의 시각적 워크플로 빌더를 쓴다.