AI 뉴스와 용어
월·수·금 오전 7시

METAL
First, AI Magazine

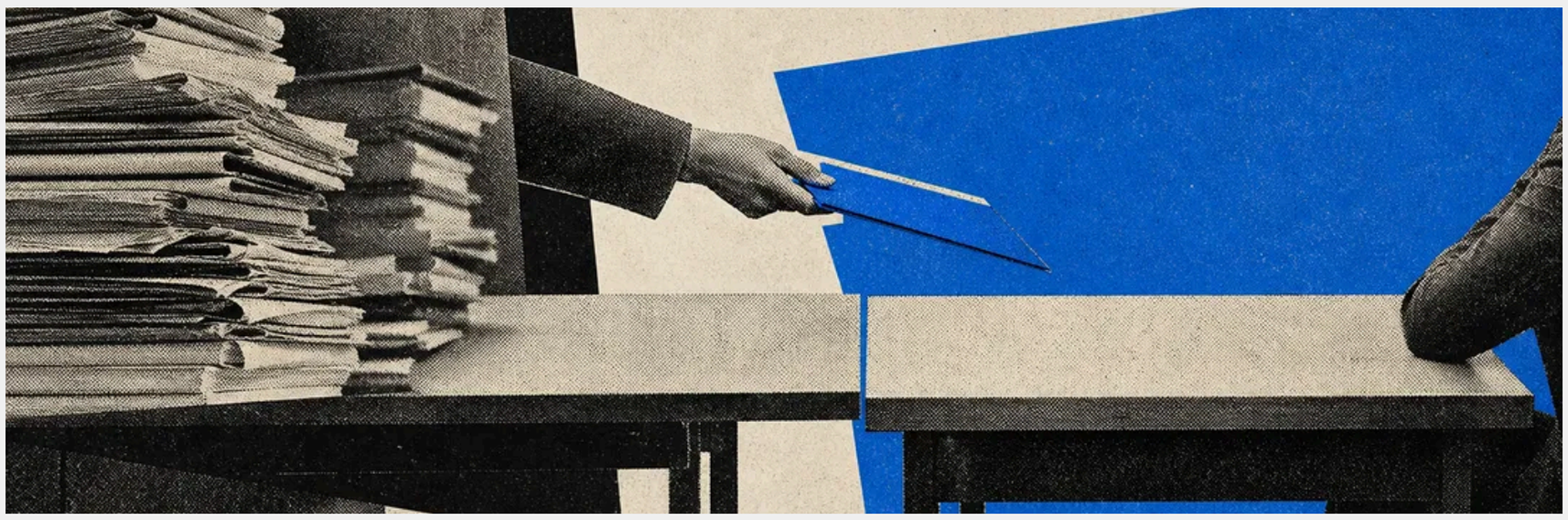
무료 · 언제든 수신거부
metallab.ai

제 22 호 | 2026년 9월 11일 금요일 | 오늘의 AI

AI

딥시크, 1.6조 모델 접고 552B로 갈아탄다

한 달 전 낸 V4-Pro를 9월 14일부터 새 모델로 넘겨요. 백본은 3분의 1인데 KV 캐시를 토큰당 890바이트로 줄여 코딩 에이전트 평가에서 오퍼스 5를 앞섰어요.



AI · 딥시크, 1.6조 모델 접고 552B로 갈아탄다

AI 모델을 만드는 딥시크가 새 모델 V4.1-Flash를 MIT 라이선스로 공개하면서, 한 달 전에 정식 출시한 플래그십 V4-Pro를 접기로 했어요. V4-Pro는 백본 파라미터가 1조 6천억 개인데 새 모델은 5,520억 개로 3분의 1 수준이에요. 회사는 허깅페이스 모델 카드에 가중치와 기술 리포트를 함께 올렸고, 9월 14일 베이징시간 정오부터 V4-Pro로 들어오는 요청을 모두 새 모델로 넘긴다고 API 문서에 적었어요.

기술 리포트의 제목이 이 모델의 정체를 말해 줘요. KV 캐시 압축의 한계를 밀어붙였다는 제목이에요. KV 캐시는 모델이 앞서 읽은 내용을 기억해 두는 임시 저장 공간인데, 문맥이 길어질수록 메모리를 잡아먹어서 비용과 속도를 좌우해요. 딥시크는 이 저장 공간을 토큰 하나당 890바이트까지 줄였다고 밝혔어요. 직전 세대인 V4-Flash의 약 4분의 1이고, 첫 세대 V1과 비교하면 437분의 1이에요.

줄인 방법은 모델을 두 토막으로 나눈 데 있어요. 회사는 이 구조를 인과 인코더 디코더라고 부르고, 40개 층을 앞의 20개 인코더와 뒤의 20개 디코더로 갈랐어요. 회사는 "디코더의 글로벌 KV 캐시가 각 디코더 층의 자체 은닉 상태에서 나오는 대신 최종 인코더 은닉 상태에서 투영됩니다"라고 적었어요. 층마다 따로 기억을 쌓지 않고, 앞쪽이 정리해 넘긴 것을 뒤쪽이 받아 쓰는 셈이에요.

앞 두 층만 슬라이딩 윈도 어텐션이고 나머지는 전부 CSA2예요.

절약이 실제로 어디에 떨어지는지도 리포트가 갈라 놓았어요. 890바이트짜리 글로벌 KV 캐시는 항상 HBM에 올라가 있고 이게 V4-Flash의 약 4분의 1이에요. 여기에 SWA 제한 재생이라는 배포 최적화를 더해, SSD나 호스트 메모리에 상주하는 KV 캐시는 V4-Flash의 약 8분의 1로 줄었다고 회사는 설명했어요. 그 8분의 1은 두 가지가 곱해진 결과예요. 상주 캐시가 SWA KV를 더 이상 담지 않아 크기가 거의 절반이 되고, 남아 있는 글로벌 KV가 다시 4분의 1로 압축돼요. 이전 세대 배포에서는 SWA KV가 상주 캐시 용량의 절반 가까이를 차지했어요.

이 설계가 만드는 차이는 활성 파라미터 수에서 드러나요. 입력을 읽어 들이는 단계에서는 토큰당 80억 개만, 답을 만들어 내는 단계에서는 160억 개만 켜져요. 회사는 이 비대칭이 입력이 많은 에이전트 작업의 비용 효율을 크게 높인다고 설명했어요. 문맥은 100만 토큰까지 받고, 이미지도 텍스트와 함께 처음부터 같이 학습해 바로 처리해요.

연산 쪽에 붙인 확장 장치도 둘 더 있어요. Engram은 기억을 계산에서 떼어 내려고 만든 조건부 메모리 모듈인데, 파라미터 1,960억개를 두 모듈에 나눠 담았어요. 반복되는 패턴을 매번 신경망으로 계산하는 대신 조회로 처리하겠다는 구조예요. DSpark는 투기적 디코딩 쪽이에요. 초안 생성기가 트랜스포머 블록 세 개에 128토큰짜리 슬라이딩 창을 써서 한 번의 순전파로 다섯 자리의 초안을 동시에 계산하고, 신뢰도 예측 헤드가 자리마다 수용 확률을 매겨요. 스케줄러는 그 값을 실측 처리량 곡선과 합쳐 요청마다 검증 길이를 그때그때 정해요.

51.8에 20포인트 넘게 못 미쳤고, 전문가 문제를 모은 HLE는 36.8로 오퍼스 5의 56.3과 격차가 커요. 모델 자체에 든 지식을 재는 SimpleQA-Verified에서는 42.3으로 V4-Pro의 55.2보다 낮아요. 백본이 3분의 1이니 외워 둔 지식에서는 큰 모델이 앞서는 셈이에요.

모델 카드에는 이 숫자들을 어떻게 읽어야 하는지 알려 주는 표가 하나 더 있어요. 같은 모델을 여덟 가지 에이전트 껍데기에 얹어 DeepSWE v1.1을 돌린 결과인데, 점수가 65.6에서 74.2까지 갈렸어요. 딥시크가 발표에 쓴 74.2는 mini-SWE 기준이고, 클로드 코드로 돌리면 69.8이에요. 메탈은 딥시크가 클로드 코드를 겨냥해 하니스팀을 새로 꾸린 일을 보도한 바 있는데, 그 팀이 만든 자체 하네스 DSH Minimal은 Terminal-Bench 2.1에서 90.6으로 클로드 코드의 88.0보다 높게 나왔어요.

성능이 오른 자리를 회사는 알고리즘이 아니라 데이터에서 찾았어요. 후처리는 지도학습과 강화학습, 정책 내 증류라는 표준 순서를 그대로 따랐고 알고리즘은 손대지 않았다고 밝혔어요. 대신 에이전트 과제와 실행 환경을 대규모로 자동 합성해 데이터와 과제, 롤아웃을 단계적으로 늘렸어요. 사전학습은 멀티모달 데이터 45조 토큰으로 했고, 희소 어텐션은 6만 4천 토큰 길이로 훈련한 뒤 34조 토큰 지점에서 100만 토큰까지 늘렸어요.

V4-Pro를 접는 일정은 문서에 날짜까지 적혀 있어요. 9월 14일 베이징시간 정오부터 deepseek-v4-pro로 오는 요청은 전부 새 모델이 처리하고 요금도 새 모델 기준으로 매겨져요. 회사는 "V4.1 Flash가 성능과 비용, 속도, 총 소요 시간에서 V4-Pro를 전면적으로 넘어섰습니다"라고 적었어요. 메탈은 V4-Pro가 추론 강도를 3단계로 조절하며

오늘 지면

이 지면은 2026년 9월 11일 금요일 아침 구독자 메일함으로 나간 레터를 신문 판형으로 다시 조판한 것입니다. 기사 전문은 metallab.ai 에서 읽을 수 있습니다.

비즈니스

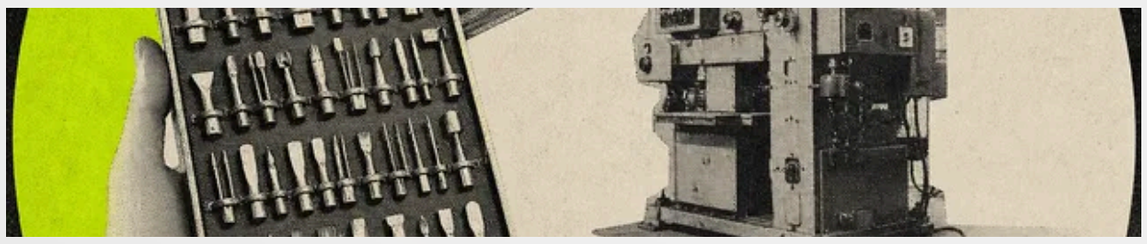
애플, 접는 아이폰과 2나노 A20 Pro 내놨다

애플이 9월 9일 이벤트에서 접히는 iPhone Duo와 2나노 공정으로 만든 A20 Pro를 함께 내놨어요. 기기 쪽 소식이 눈에 먼저 들어오지만, 이 발표에서 실제로 바뀐 축은 기기 안에서 AI를 돌리는 계산 능력이에요. A20 Pro의 뉴럴 엔진이 16코어짜리 두 개, 합쳐서 32코어가 되면서 기기 안에서 처리하는 AI 연산 능력이 이전 세대의 두 배가 됐어요. 화면에 보이는 것을 읽고 앱 사이의 맥락을 잡는 Siri AI가 담긴 iOS

▶ 2면에 계속

AI

5년 된 폰에서 도는 모델 18개 나왔다

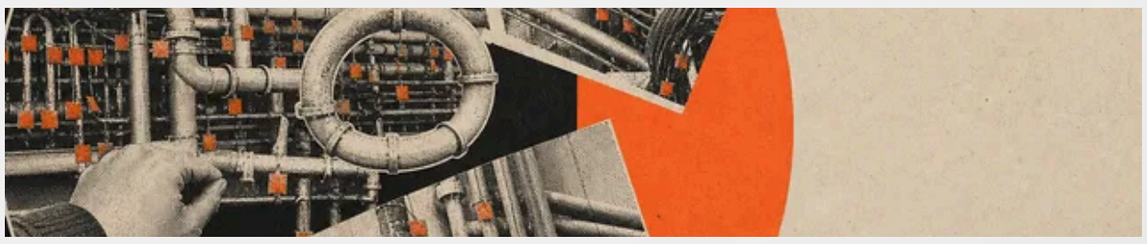


유럽에 새로 생긴 AI 랩 Desert Ant Labs가 기기 안에서 도는 모델 18개를 한꺼번에 내놨어요. 9월 8일 올린 출범 글에서 회사는 안정판 12개와 베타 6개를 Swift와 Kotlin, JavaScript용 SDK 하나로 묶었다고 밝혔어요. 큰 모델 하나가 여러 일을 맡는 대신 과제 하나에 모델 하나를 붙이는 구성이에요.

회사가 내세운 기준은 크기가 아니라 응답 시간이에요. 창업자 겸 대표 폴 뵈헌은 "우리는 오디오와 비전, 텍스트를 위한 작고 특화된 모델을 만들고 있고, 각 모델은 밀리초 단위로 답합니다"라고 적었어요. 5년 된 폰에서도 돌아가고 영상 프레임마다 또는 자판을 누를 때마다 불러도 버틸 만큼 빨라야 한다는 게 설계 조건이에요. 호출 한 번의 비용이 사라지면 확인할 수 있는 것만 확인하는 대신 모든 메시지에 기능을 걸 수 있다는 게 회사의 주장이에요.

AI

오픈AI, 250명 투입해 사내 취약점 훑었다



챗GPT를 만드는 오픈AI가 사내에 코드 레드를 걸고 자사 시스템 전체를 훑는 보안 스프린트를 진행했다고 밝혔어요. 투입된 인력은 250명이 넘고 대상은 100개가 넘는 서비스 영역이었어요. 회사는 이 작업에서 쓴 참조 아키텍처와 실행 순서를 Defense Factory라는 이름으로 공개했어요. 자기 보안 조직을 어떻게 다시 짰는지를 경쟁사까지 열어 놓고 보여 주는 문서예요.

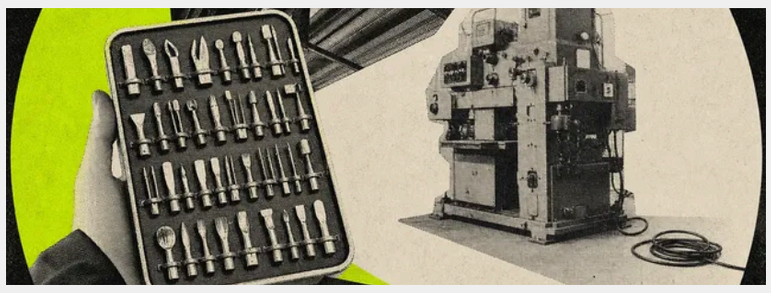
회사가 급해진 이유는 공격 쪽의 조건이 바뀌었기 때문이에요. 오픈AI는 널리 풀린 오픈 웨이트 모델을 쓰는 에이전트가 이제 오랜 시간에 걸친 사이버 작업을 수행할 수 있다고 설명했어요. 에이전트가 세션 사이에 배운 것을 남겨 두면 시스템 구조를 상세히 파악하고 서로 다른 취약점을 이어 붙일 수 있어요. 예전에는 손이 많이 가서 시도조차 어렵던 연계 공격이 이제는 자율적으로 굴러간다는 뜻이에요.

▶ 2면에 계속

◀ 1면에서 계속

AI

5년 된 폰에서 도는 모델 18개 나왔다



1 Desert Ant Labs가 9월 8일 출범하면서 기기 안에서 도는 모델 18개를 공개했어요. 안정판 12개와 베타 6개를 Swift·Kotlin·JavaScript용 SDK 하나로 묶었어요.

2 받아쓰기 모델 Voz는 아이폰에서 10분 오디오를 2초에 처리하고, 30분 연속 기준 실시간의 319배로 Whisper large-v3-turbo의 50배를 크게 앞섰어요.

3 12MB짜리 Redact는 개인정보 88.8%를 잡아 2.3GB GLiNER-PII의 91.1%에는 못 미쳤어요. 회사는 자기 영상 앱에서 Claude Sonnet을 284MB Clips로 바꿨다고 밝혔어요.

숫자는 과제별로 나뉘어 있어요. 받아쓰기 모델 Voz는 아이폰에서 10분짜리 오디오를 2초에 옮기고 낱말마다 시작과 끝 시각을 붙여요. 회사가 공개한 측정에서 30분 연속 오디오 기준 실시간 대비 배속은 Voz가 319배, 애플 SpeechAnalyzer가 78배, Whisper large-v3-turbo가 50배였어요. 아이폰 17 Pro에서는 298배가 나왔어요.

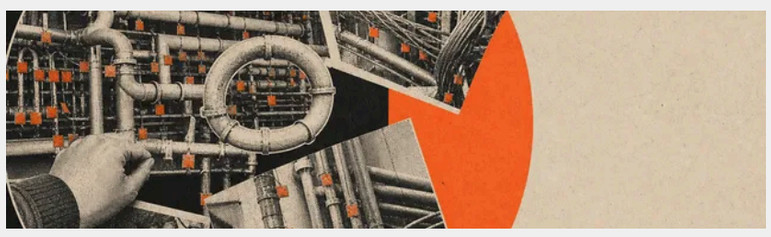
오디오를 다듬는 Clear는 9MB짜리예요. 5분 길이 노트북 녹음을 1초 만에 스튜디오 품질로 바꾼다고 회사는 설명했어요. 아이폰 16 Pro에서 실시간의 302배, M5 맥북 프로에서 345배로 측정됐어요. 같은 가중치가 아이폰에서는 뉴럴 엔진으로, 브라우저에서는 WebAssembly로 돌아가요.

개인정보를 가리는 Redact에서는 회사가 밀린 대목도 그대로 실렸어요. 12MB인 Redact는 문서에 든 개인정보의 88.8%를 잡아내는데, 2.3GB짜리 GLiNER-PII의 91.1%보다 낮아요. 대신 14.7MB의 Rampart는 61.4%, 3GB짜리 오픈AI 필터는 60.2%에 그쳤어요. 언어를 가려내는 Tongue은 2MB로 세 낱말만 보고 84개 언어를 구분해 0.933을 기록했고, 293MB짜리 판별기는 0.887이었어요.

◀ 1면에서 계속

AI

오픈AI, 250명 투입해 사내 취약점 훑었다



1 오픈AI가 사내 코드 레드를 발령하고 250명 넘는 인력으로 100개 넘는 서비스 영역의 취약점을 찾아 검증하고 고치는 보안 스프린트를 진행했다고 밝혔어요.

2 첫날 긴급·높음 이슈 53건 처리, 발견 사항의 37%가 중복, 19.5%가 런타임에서 재현, 동적 검증 뒤 오탐률 0.81%, 담당자 배정 수락률 90.6%, 롤백률 0.53%를 공개했어요.

3 널리 풀린 오픈 웨이트 모델을 쓰는 에이전트가 취약점을 이어 붙여 자율 공격을 할 수 있게 되면서, 회사는 방어자에게 남은 시차가 닫히기 전에 움직여야 한다고 주장했어요.

그래도 방어 쪽에 아직 시간이 남아 있다는 게 회사의 주장이에요. 오픈AI는 방어자의 구조적 이점을 두고 "에이전트가 자사 코드에 직접 액세스할 수 있도록 하고, 프런티어 모델을 활용해 널리 이용 가능한 오픈 웨이트 모델을 악용하는 공격자보다 먼저 앞서 나갈 수 있습니다"라고 적었어요. 자기 코드를 통째로 열어 줄 수 있는 쪽도, 최신 모델을 먼저 쓸 수 있는 쪽도 방어자뿐이라는 거예요. 회사는 이 시차를 방어자에게 열린 기회라고 부르면서 지금 움직이지 않으면 닫힌다고 못 박았어요.

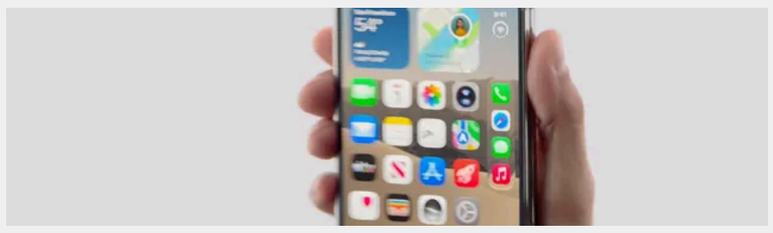
스프린트는 실제 사고가 난 것처럼 굴러갔어요. 보안과 응용, 연구 조직을 한자리에 모아 수백 개 시스템을 한꺼번에 걸었어요. 오픈AI 코어 제품·플랫폼 총괄 티보는 "OpenAI는 인시던트에 대응할 때와 같은 긴박감을 가지고 방어 체계를 강화하고 있습니다"라고 말했어요. 그는 이 작업이 핵심 사업 운영을 뺀 다른 모든 업무보다 앞서는 전사 과제이며 스프린트가 끝난 뒤에도 같은 긴박감을 유지하겠다고 덧붙였어요. 첫날에만 긴급 또는 높은 우선순위 이슈 53건을 처리했어요.

메탈이 확인한 공개 문서에는 잘 굴러가지 않았던 대목의 숫자도 그대로 적혀 있어요. 처음에는 심각도 분류의 범주가 너무 넓어서 에이전트에 준 지침에 따라 결과가 흔들렸어요. 회사는 평가 기준과 프롬프트에 버전을 매기고 검토자가 기대한 우선순위와 판단 근거를 기록한 뒤에야 일괄 처리로 넘어갔어요. 그 과정에서 발견 사항의 37%가 중복으로 확인됐고, 중복 제거 성능이 나아질 때까지 담당자 배정을 아예 멈춰 뒀어요.

◀ 1면에서 계속

비즈니스

애플, 접는 아이폰과 2나노 A20 Pro 내놨다



1 애플이 9월 9일 첫 접이식 iPhone Duo와 2나노 A20 Pro를 공개했어요. 뉴럴 엔진이 16코어 두 벌 합쳐 32코어가 되면서 기기 안 AI 연산 능력이 두 배가 됐어요.

2 A20 Pro는 6코어 CPU에 데스크톱급 슈퍼 코어 두 개를 넣어 A19 Pro보다 20% 빠르고, 7코어 GPU는 40% 빨라요. 메모리 대역폭은 이전 세대보다 50% 넓어졌어요.

3 Siri AI를 담은 iOS 27과 macOS 27 골든게이트는 9월 14일에 나오고, iPhone Duo는 10월 23일 70개 넘는 나라에서 1,999달러부터 팔려요.

기기 안에서 모델을 돌리는 쪽에서 이 숫자가 왜 중요한지는 분명해요. 연산 능력만큼 먼저 걸리는 것이 발열과 대역폭인데, 이번에는 두 가지를 같이 손봤어요. 메모리 대역폭이 넓어지면 큰 모델의 가중치를 읽어 오는 구간에서 병목이 줄어요. 열 쪽도 실리콘을 증기 챔버에 바로 붙이는 구조로 바꿨고, 함께 나온 iPhone 18 Pro는 이 냉각 설계로 iPhone 17 Pro보다 지속 성능이 40% 빠르다고 회사는 밝혔어요. 화면을 계속 읽고 앱 사이를 오가며 맥락을 유지하는 비서는 짧게 튀는 성능이 아니라 이쪽을 필요로 해요.

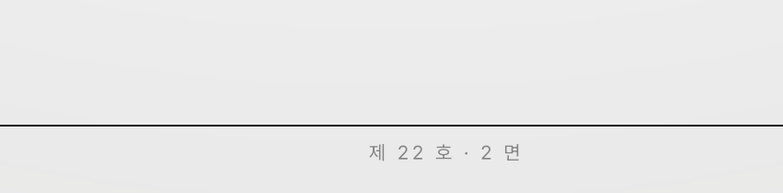
iPhone Duo의 규격도 정리해 둘게요. 펼치면 7.6인치 화면이고, 접으면 5.4인치 바깥 화면이 iPhone 18 Pro 화면 면적의 90%를 담아요. 두 화면 모두 프로모션과 상시 표시를 지원하고 최대 밝기는 3,000니트예요. 나노텍스처 마감을 넣었고, 몸체는 거울처럼 다듬은 5등급 티타늄이에요. 앞은 긁힘에 세 배 강해졌다는 세라믹 실드 2, 뒤는 세라믹 실드로 덮었어요. 경첩은 100개가 넘는 부품으로 짰고, 방수와 방진은 IP68로 6미터에서 30분을 견뎌요.

배터리는 양쪽에 하나씩 두 개를 넣었어요. 안쪽 화면으로 영상을 볼 때 최대 31시간, 바깥 화면으로는 44시간, 섞어 쓰면 한 번 충전에 24시간이에요. 유선으로 20분, 무선으로 30분이면 절반까지 찬다고 회사는 설명했어요. 카메라는 4,800만 화소 메인과 같은 화소의 초광각이고, 2배 광학 망원과 4K 120프레임 돌비 비전 녹화를 지원해요. 화상 통화용 카메라는 화면 아래에 숨겼어요. 값은 256기가 기준 1,999달러이고, 예약은 10월 16일, 판매는 10월 23일에 70개 넘는 나라에서 시작해요. 최고경영자 존 터너스는 iPhone Duo를 두고 "초대 iPhone 이후 가장 큰 변화"라고 했어요.

AI

오픈AI, 회사 데이터 묻는 에이전트 내놨다

스노우플레이크나 빅쿼리에 직접 붙어 무엇이 달라졌는지 조사하고 대시보드까지 만들어 줘요. 쿼리를 쓸 줄 몰라도 되고, 계정에 걸린 행과 열 권한은 그대로 지켜져요.



1 오픈AI가 9월 10일 ChatGPT Work에 Data agent를 넣었어요. Amazon Redshift와 Google BigQuery, ClickHouse, Databricks, MongoDB, Snowflake 등에 붙어 대화만으로 분석하고 대시보드를 만들어요.

2 회사 용어와 지표 정의는 Databricks Genie Ontology, dbt, GitHub, Snowflake Horizon 같은 의미 계층에서 읽어 오고, 질의는 연결된 계정의 테이블·행·열 권한을 그대로 따라요.

3 오픈AI는 제품 조직 거의 전원과 GTM 조직 3분의 2 이상이 사내에서 이미 쓰고 있다고 밝혔어요. 사례로 실린 고객 열한 곳은 모두 알파 단계 참가사이거나 평가 파트너예요.

오픈AI가 회사 데이터에 말로 묻고 대시보드까지 만들어 주는 도구를 ChatGPT Work에 넣었어요. 9월 10일 공개한 Data agent는 사내 데이터베이스에 직접 붙어 무엇이 달라졌는지 조사하고, 공유할 수 있는 대화형 대시보드를 만들어요. 회사는 쿼리를 쓰거나 새 분석 도구를 배우지 않아도 대화 하나로 분석을 지시하고 다듬을 수 있다고 설명했어요.

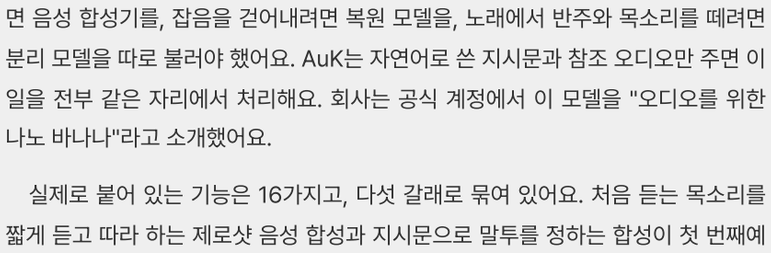
붙는 곳이 꽤 넓어요. 승인된 데이터 소스로 Amazon Redshift와 Datadog, Google BigQuery, ClickHouse, Databricks, MongoDB, Snowflake 등을 들었고, Google Drive와 SharePoint에 있는 파일과 문서도 분석에 끌어올 수 있어요. 결과를 놓는 자리도 하나가 아니에요. Omni와 Oracle BI, Power BI, Sigma, Tableau, ThoughtSpot 안에서 대시보드를 만들고 손볼 수 있어요.

회사 안에서만 통하는 말을 어떻게 알아듣느냐가 이런 도구의 승부처예요. 오픈AI는 조직의 업무 용어와 지표 정의, 맞춤 계산식, 데이터 사이의 관계를 읽어 해석한다고 밝혔어요. 그 맥락은 Databricks Genie Ontology와 dbt, GitHub, Snowflake Horizon, BI 대시보드 같은 곳에 이미 쌓여 있는 의미 계층에서 가져와요. 활성 고객이라는 말이 회사마다 다른 뜻인데, 그 정의를 새로 받아 적는 대신 이미 정해 둔 곳을 읽는 방식이에요.

AI

텐센트, 목소리 만들고 고치는 모델 풀었다

자연어로 지시하면 목소리를 만들고 감정과 말투까지 바꿔 줘요. 코드와 가중치를 MIT 라이선스로 열었고, 네 단계 만에 답하는 경량판도 같이 나왔어요.



1 텐센트 훈위안이 음성 생성과 편집을 한 모델에 합친 AuK를 오픈소스로 공개했어요. 파라미터는 15억개, 라이선스는 MIT이고 코드와 가중치를 함께 열었어요.

2 자연어 지시문과 참조 오디오라는 입구 하나로 제로샷 합성, 내용 편집, 감정과 음색 변경, 잡음 제거와 음악 분리까지 16가지 과제를 처리해요.

3 학습에는 지시문과 오디오를 짝지은 사례 30억 3천만건과 감독 데이터 195만 시간이 들어갔고, 경량판 AuK-Flash는 네 단계 추론으로 약 4.5배 빨랐어요.

텐센트 훈위안이 목소리를 만드는 일과 고치는 일을 하나로 묶은 모델 AuK를 오픈소스로 공개했어요. 코드와 가중치를 깃허브 저장소에 함께 올렸고 라이선스는 MIT예요. 파라미터는 15억개로, 요즘 나오는 대형 모델과 견주면 작은 축에 들어요.

이번 공개에서 눈여겨볼 대목은 성능 순위가 아니라 입구를 하나로 줄였다는 데 있어요. 지금까지 오디오 작업은 도구를 갈아 끼우는 일이었어요. 문장을 소리로 읽히려면 음성 합성기를, 잡음을 걷어내려면 복원 모델을, 노래에서 반주와 목소리를 떼려면 분리 모델을 따로 불러야 했어요. AuK는 자연어로 쓴 지시문과 참조 오디오만 주면 이 일을 전부 같은 자리에서 처리해요. 회사는 공식 계정에서 이 모델을 "오디오를 위한 나노 바나나"라고 소개했어요.

실제로 붙어 있는 기능은 16가지고, 다섯 갈래로 묶여 있어요. 처음 듣는 목소리를 짧게 듣고 따라 하는 제로샷 음성 합성과 지시문으로 말투를 정하는 합성이 첫 번째예요. 녹음된 말의 내용을 바꿔 끼우거나 가사를 고치는 내용 편집이 두 번째고, 높낮이와 속도와 크기를 조절하는 음향 편집이 세 번째예요. 감정과 음색을 바꾸고 억양을 지우고 말이 아닌 소리를 걷어내고 보통 말을 속삭임으로 바꾸는 준언어 편집이 네 번째, 잡음 제거와 화자 분리와 음악 분리와 특정 화자 추출이 다섯 번째예요.