AI 뉴스와 용어
월·수·금 오전 7시

METAL
First, AI Magazine

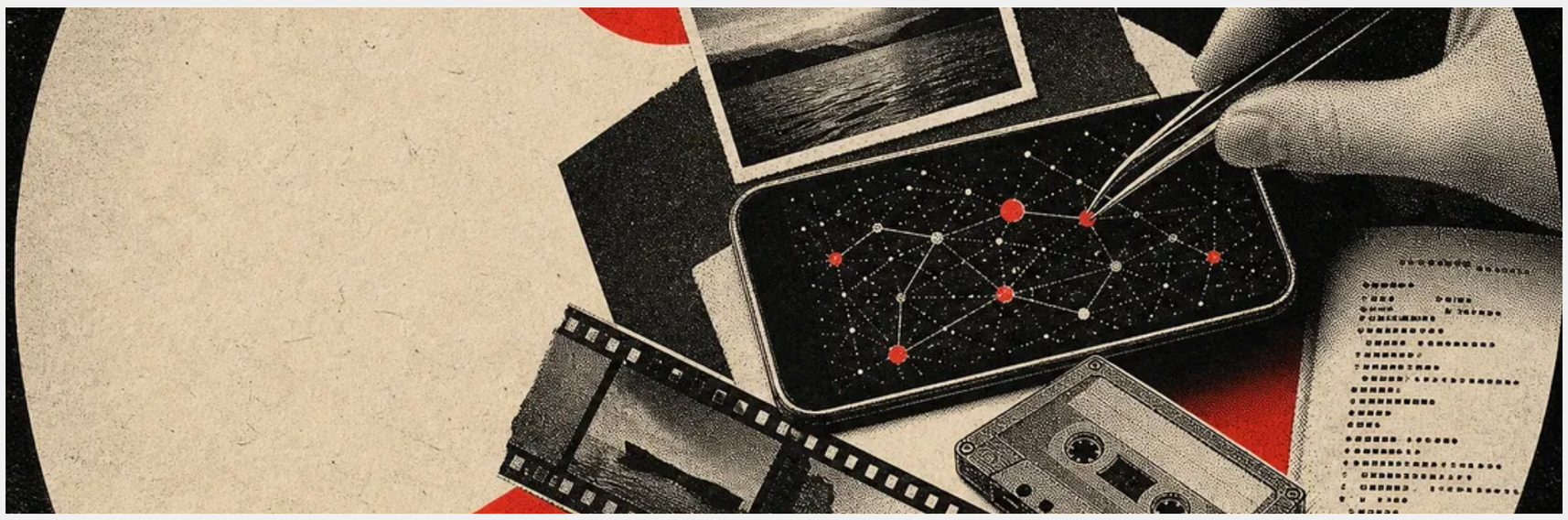
무료 · 언제든 수신거부
metallab.ai

제 40 호 | 2026년 10월 8일 목요일 | 오늘의 AI

AI

# 구글 딥마인드, EmbeddingGemma 2 공개

구글 딥마인드가 텍스트와 코드, 이미지, 오디오, 영상을 한 임베딩 공간에 담는 7억4000만 파라미터 공개 모델 EmbeddingGemma 2를 내놓았다. 필요한 인코더만 골라 얹는 조립식 구조로 스마트폰에서 멀티모달 검색을 돌린다.

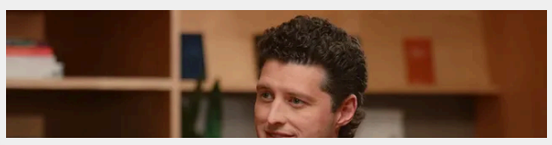


AI · 구글 딥마인드, EmbeddingGemma 2 공개

구글 딥마인드가 10월 6일 텍스트와 코드, 이미지, 오디오, 영상을 하나의 임베딩 공간에 담는 공개 모델 EmbeddingGemma 2를 내놓았다. 파라미터는 7억4000만 개로, 스마트폰과 노트북 같은 개인 기기 안에서 돌아가도록 설계됐다. 회사는 이 모델을 온디바이스 멀티모달 임베딩에서 가장 뛰어난 모델이라고 소개했다. 상업적 이용을 허용하는 아파치 2.0 라이선스로 풀렸고, 가중치는 허깅페이스와 캐글에서 받을 수 있다.

임베딩 모델은 글이나 사진, 소리를 숫자 목록인 벡터로 바꿔 뜻이 가까운 것끼리 가까이 놓는다. 검색과 추천, 검색 증강 생성(RAG)의 밑바닥에 깔리는 부품이다. 지난해 9월 나온 1세대 EmbeddingGemma는 텍스트만 다뤘다. 발표문을 쓴 딥마인드 리서치 엔지니어 사힐 두아와 엔리케 셰크터 베라는 1세대가 2000만 회 넘게 내려받혔다며 "개발자 커뮤니티의 반응은 우리의 기대를 훨씬 뛰어넘었다"고 밝혔다. 개발자들은 이 모델로 기기 안 검색 도구와 개인정보를 밖으로 보내지 않는 RAG 파이프라인을 만들었다.

2세대의 핵심은 감각을 하나로 묶은 것이다. 두 엔지니어는 이 모델이 "음성 메모로 특정 영상 장면을 찾거나, 텍스트 질문으로 몇 시간 분량의 오디오 녹음을 뒤질 수 있고, 이 모든 처리를 하나의 네이티브 멀티모달 모델이 한다"고 설명했다. 기반은 올해 4월 공개된 Gemma 4 아키텍처다. 구글은 이 모델이 제미나이 임베딩 모델과 같은 기술로 만들어졌다고 밝혔다.



구조는 조립식이다. 텍스트와 코드만 처리할 때는 2억7000만 개 파라미터면 되고, 1억7000만 개짜리 비전 인코더와 3억 개짜리 오디오 인코더를 필요할 때만 붙인다. 보도에 따르면 모델 카드는 텍스트와 비전 조합을 4억4000만 개, 텍스트와 오디오 조합을 5억7000만 개로 적고 있으며, 어떤 조합으로 불러와도 같은 768차원 벡터 공간을 쓴다. 사진만 검색하는 앱이 오디오 인코더까지 메모리에 올릴 필요가 없다는 뜻이다.

메모리 수치도 구체적이다. 구글은 양자화한 모델을 픽셀 11 프로에서 돌렸을 때 텍스트 전용 가중치는 활성 RAM 약 191MB, 멀티모달 전체는 약 567MB를 쓴다고 밝혔다. 문맥 창은 8192토큰으로 1세대의 네 배다. 한 번에 오디오 5.5분이나 이미지 29장, 영상 프레임 58장, 또는 이를 섞은 입력을 넣을 수 있다. 보도에 따르면 이미지 한 장은 280토큰, 영상 프레임 하나는 140토큰, 오디오 1초는 25토큰을 차지한다.

저장 공간을 줄이는 장치도 넣었다. 마트료시카 표현 학습(MRL)으로 768차원 출력 벡터를 512, 256, 128차원으로 잘라 쓸 수 있고, 회사는 이렇게 하면 로컬 벡터 데이터베이스의 저장 공간과 메모리를 최대 6배 줄인다고 설명했다. 보도에 따르면 구글 개발자 가이드는 256차원에서 이미지·영상·음성 검색 품질이 원래의 약 95%로 유지되고, 128차원에서는 텍스트와 코드가 약 90%를 지키지만 멀티모달 검색은 약 75%로 떨어진다고 적었다.

성적표에서 가장 크게 오른 것은 코드다. 대규모 텍스트 임베딩 벤치마크(MTEB) 코드 부문 점수가 68.76에서 78.68로 9.92점 올랐고, 다국어 텍스트 성능은 1세대 수준을 유지했다. 메탈이 확인한 구글 발표문의 MTEB 코드 점수를 냈고, 이미지와 영상, 오디오 과제에서 두 배 넘게 큰 전문 모델 일부를 앞섰다고 밝혔다.

엔지니어 관점에서 눈여겨볼 대목은 Gemma 4와 부품을 나눠 쓴다는 점이다. EmbeddingGemma 2는 Gemma 4의 텍스트 토크나이저와 오디오 인코더를 공유해, 두 모델을 한 파이프라인으로 돌리면 전체 메모리 사용량이 줄어든다. 찾는 모델과 답하는 모델이 한 기기 안에서 같은 부품을 쓰는 구조다. 구글은 AI 엣지 갤러리 앱에 영상 속 장면을 텍스트나 음성으로 찾는 비디오 모먼츠 파인더 데모를 실었다. 맥용 회의 앱 AI 엣지 포어사이트는 이 모델로 로컬 파일을 찾고 Gemma 4로 맥락을 추론한다.

배포 경로도 넓게 깔았다. 트랜스포머스와 센턴스 트랜스포머스, MLX, vLLM, llama.cpp, SGLang, 올라마, LM스튜디오에서 돌릴 수 있고, 미세조정은 언슬로스 가이드를 따른다. 기기용으로는 미디어파이프와 LiteRT를, 브라우저용으로는 transformers.js와 WebGPU를 지원한다. 제미나이 엔터프라이즈 에이전트 플랫폼 모델 가든 탑재는 곧 이뤄진다고 회사는 밝혔다. 보도에 따르면 모델 카드는 활성값 범위가 float16을 넘어 NaN이 생길 수 있으니 bfloat16이나 float32로 추론하라고 안내하고, 학습 데이터 기준 시점은 2025년 1월이다.

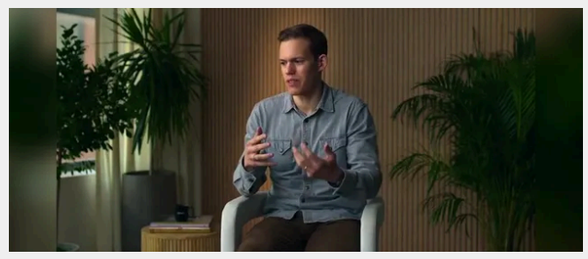
메탈은 리퀴드AI가 온디바이스 모델 평가 도구 Pipette를 오픈소스로 공개했다고 보도한 바 있다. 온디바이스 경쟁이 생성 모델에서 검색 부품으로 내려오고 있다. 서버에 사진과 녹음을 올리지 않고도 기기 안에서 영상 한 장면을 찾아내는 일이 메모리 600MB 아래의 부품으로 가능해졌다. 이제 앱 개발자가 정할 것은 어떤 인코더를 얹고 벡터를 몇 차원까지 줄일지다.

오늘 지면

1 AI 구글 딥마인드, EmbeddingGemma 2 공개
2 AI 오픈AI, 수학 논문 722편 깃허브 공개
3 AI 오픈AI, Decisions API 베타 공개
4 AI 오픈AI, 점프 트레이딩 Astra 사례 공개
5 AI 구글, SynthID Detector 일반 공개
6 AI 에브리, 슬랙용 에브리 에이전트 공개

이 지면은 2026년 10월 8일 목요일 아침 구독자 메일함으로 나간 레터를 신문 판형으로 다시 조판한 것입니다. 기사 전문은 metallab.ai 에서 읽을 수 있습니다.

AI



## 오픈AI, 점프 트레이딩 Astra 사례 공개

북미 퀀트 트레이딩 회사 점프 트레이딩(Jump Trading)이 오픈AI의 GPT-6 Astra로 며칠씩 돌아가는 시장 연구를 에이전트에게 맡기고 있다. 오픈AI는 10월 6일 고객 사례를 내고 점프 트레이딩이 챗GPT와 GPT-6 Astra로 더 길고 모호한 연구 문제를 다루고 있다고 밝혔다. 회사에서 LLM 연구개발을 이끄는 루카스 베이커 책임자는 "GPT-6 시리즈, 특히 GPT-6 Astra로 오픈AI는 장기 과제에서 새로운 단계의 자

▶ 2면에 계속

AI

## 오픈AI, 수학 논문 722편 깃허브 공개



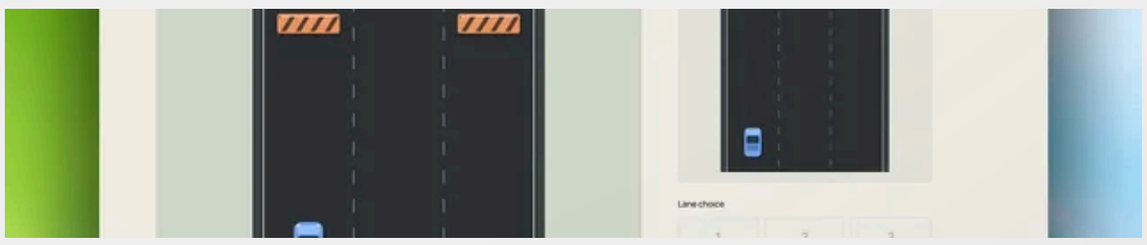
오픈AI가 10월 6일 아직 공개하지 않은 내부 프런티어 모델이 만든 수학 논문 722편을 깃허브 저장소 openai/math에 한꺼번에 올렸다. 논문은 관련 결과끼리 묶은 372개 결과군(family)으로 정리됐고, 상당수에는 컴퓨터가 증명을 검사할 수 있는 린(Lean) 형식화가 붙었다. 회사는 모델이 이번 평가 기간에 약 4000개 문제를 받았고, 결과 하나에 평균 챗GPT 프로의 사고 연산 약 3시간에 해당하는 계산이 들었다고 밝혔다. 형식화가 없는 결과에는 문제가 있을 수 있다는 단서도 저장소에 함께 적었다.

메탈이 확인한 저장소의 41쪽짜리 개요 PDF에는 정수론과 대수기하, 이론 컴퓨터과학, 조합론, 작용소 대수, 편미분방정식 등 17개 분야로 나눈 목록이 실려 있다. 목록이 내건

▶ 2면에 계속

AI

## 오픈AI, Decisions API 베타 공개




오픈AI가 질문 하나에 정해진 형식의 답만 돌려주는 Decisions API를 10월 6일(미국 시간) 모든 개발자에게 퍼블릭 베타로 공개했다. 회사는 공식 개발자 계정에서 이 API가 "앱이 적절한 모델이나 도구, 행동을 거의 실시간으로 고르게 해 준다"고 밝혔고, Responses API로 GPT-6 Luna를 부를 때보다 최대 10배 빠르게 결정을 내린다고 설명했다. 지원 모델은 GPT-6 Luna 하나이고, 호출 주소는 전용 엔드포인트인 POST /v1/decisions다. 오픈AI는 개발자 문서에서 몇 주 안에 정식 출시(GA)할 것으로 예상한다고 적었다.

이번 공개는 9월 29일 데브데이 2026에서 제한 프리뷰로 나온 뒤 일주일 만이다. 메탈은 오픈AI가 결정 전용 Decisions API를 제한 프리뷰로 공개했다고 보도한 바 있으

▶ 2면에 계속

◀ 1면에서 계속

AI

## 오픈AI, 수학 논문 722편 깃허브 공개



1 오픈AI가 10월 6일 미공개 내부 모델이 만든 수학 논문 722편을 372개 결과군으로 묶어 깃허브 저장소 openai/math에 공개했다.

2 모델은 약 4000개 문제를 받았고 결과당 평균 챗GPT 프로 사고 연산 약 3시간이 들었으며, 상당수에 린 형식화가 붙었지만 형식화되지 않은 결과는 문제가 있을 수 있다고 회사가 밝혔다.

3 프린스턴 고등연구소 자문그룹 권고를 참고했다지만 모델 이름과 결과별 프롬프트는 공개하지 않았고, 회사는 워크숍·학회 지원과 모델 공개를 예고했다.

만든 방식도 공개했다. README에 따르면 결과 대부분은 같은 절차로 내부 모델 하나에서 나왔다. 기존 수학 평가에서 성능이 포화하자 회사가 미해결 연구 문제로 평가를 넓힌 결과이고, 일부는 모델이 앞서 낸 결과 위에 쌓였다. 예외는 리만 제타 함수의 영점 없는 영역 연구와 CM 아벨 다양체의 호지 추측 증명 두 건이며, 실수부 11/12 영역에 대한 원고는 읽기 쉽게 사람이 손봤다. 저장소에는 파이의 무리성 지수와 말러 추측, 스핀 글라스의 메자르-파리시 공식 등 10개 결과에 대한 모델 추론 요약도 실렸다.

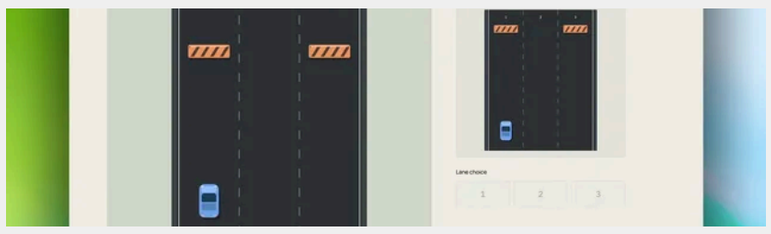
발표 방식은 외부 권고를 의식했다. 오픈AI는 프린스턴 고등연구소에 꾸려진 독립 기구 수학·인공지능 자문그룹과 상의해 공개 원칙을 정했다고 밝혔다. 메탈은 오픈AI가 이 자문그룹 출범을 발표했다고 보도한 바 있다. 보도에 따르면 자문그룹은 9월 29일 수학계 응답 600건 이상을 바탕으로 권고안을 냈고, 결과마다 모델 이름과 프롬프트, 추론 요약, 걸린 시간, 계산 비용을 공개하고 AI 연구소가 통제하지 않는 학술 저장소에 맡기라고 요구했다. 권고안은 공개되지 않은 모델로 고난도 수학 문제를 시험하는 관행에 대해 "우리는 이 관행을 지지하지 않으며, 이를 멈출 것을 요청한다"고 적었다.

이번 공개는 그 요구 가운데 일부만 따랐다. 추론 요약과 시도한 문제 수, 계산량 추정, 이전 판을 남기는 개정 정책은 내놨다. 모델 이름과 결과별 프롬프트는 빠졌고, 비용은 결과별이 아닌 평균으로만 밝혔다. 저장소도 오픈AI가 관리하는 깃허브이며, 회사는 자문그룹 기준에 맞는 커뮤니티 운영 저장소를 계속 찾겠다고 설명했다. 보도에 따르면 오픈AI는 자문그룹이 결과를 어떻게 알릴지는 조언할 수 있어도 결과를 낼지, 얼마나 빨리 낼지에는 관여하지 못하도록 선을 그었다.

◀ 1면에서 계속

AI

## 오픈AI, Decisions API 베타 공개



1 오픈AI가 10월 6일(미국 시간) Decisions API를 모든 개발자에게 퍼블릭 베타로 공개하고, Responses API로 GPT-6 Luna를 부를 때보다 최대 10배 빠르다고 밝혔다.

2 질문 유형은 조건의 참 확률을 내는 프레디킷, 선택지 하나를 고르는 초이스, 기준표 점수를 내는 스코어 세 가지이고 텍스트와 이미지를 받는다.

3 요금은 입력 토큰 100만 개당 0.10달러로 출력·캐시 토큰은 과금하지 않으며, 자격 고객에게 ZDR과 HIPAA 용도를 지원한다.

요청은 세 칸으로 이뤄진다. model 칸에는 평가할 모델을, input 칸에는 질문들이 함께 참고할 증거를 텍스트 문자열이나 텍스트·이미지가 섞인 사용자 메시지로 넣는다. questions 칸에는 질문마다 유형과 지시문, 허용할 선택지나 점수 단계를 적는다. 응답은 answers 배열로 돌아오고, 질문마다 붙인 고유 이름이 답에 그대로 되돌아와 어느 질문의 답인지 가린다.

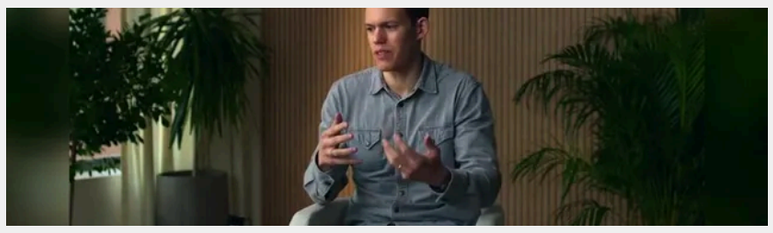
질문 유형은 세 가지다. 프레디킷(predicate)은 어떤 조건이 참일 확률을 0에서 1 사이로 추정한다. 초이스(choice)는 개발자가 준 선택지 중 하나를 고르고 선택지마다의 확률과 별도의 신뢰도 값을 함께 낸다. 스코어(score)는 낮은 단계부터 높은 단계까지 정한 기준표에 입력을 대 보고, 단계 번호의 확률 가중 평균을 점수로 돌려준다. 그래서 점수는 단계와 단계 사이에 떨어질 수 있다.

메탈이 확인한 개발자 가이드의 예시는 이 구조를 숫자로 보여 준다. 제품 사진에서 금이나 찢김, 찌그러짐을 찾는 프레디킷 질문에는 확률 0.92가 돌아왔다. 「주문 금액이 두 번 청구됐다」는 고객 불만을 부서로 나누는 초이스 질문은 결제 부서를 0.95 확률로 골랐고 신뢰도는 0.93이었다. 사파리에서만 내보내기가 실패하는 문제를 세 단계 심각도에 대 본 스코어 질문은 0.1, 0.7, 0.2의 확률 분포에서 점수 1.1, 신뢰도 0.55를 냈다. 오픈AI는 이 값들이 설명용 응답 예시라고 밝혔다.

◀ 1면에서 계속

AI

## 오픈AI, 점프 트레이딩 Astra 사례 공개



1 오픈AI가 10월 6일 점프 트레이딩이 챗GPT와 GPT-6 Astra로 길고 모호한 퀀트 연구를 에이전트에게 맡기는 사례를 공개했다.

2 점프 트레이딩 LLM 연구개발 책임자 루카스 베이커는 에이전트가 며칠씩 여러 데이터를 엮어 성과를 쌓는 재귀적 개선이 가능해졌다고 밝혔다.

3 에이전트가 낸 거래 신호도 다른 신호처럼 범위를 정해 검토하고, 사람이 마지막 검증을 맡는 구조로 규제 업종의 위험을 다룬다.

규모는 데이터에서 드러난다. 베이커가 동료 로렌 푸찰라 피오레와 함께 올해 4월 25일 ICLR 2026 엑스포 세션에서 발표한 초록에 따르면, 점프 트레이딩은 매일 테라바이트 단위의 시장 데이터를 처리하며 전 세계 수천 개 종목에서 예측 신호를 찾는다. 파이프라인은 호가창의 원시 이벤트부터 자연어 신호 통합, 모델 학습, 실거래 집행까지 이어지고, 모두 엄격한 지연 시간 제약 아래 돈다. 회사는 이 자리에서 텍스트 입력으로 대형 언어 모델을 미세 조정해 실거래 신호를 만든 결과와, 비정형 자료를 묶어 사람 트레이더와 자동 전략이 함께 쓰는 예측을 내는 다중 에이전트 시스템을 소개했다.

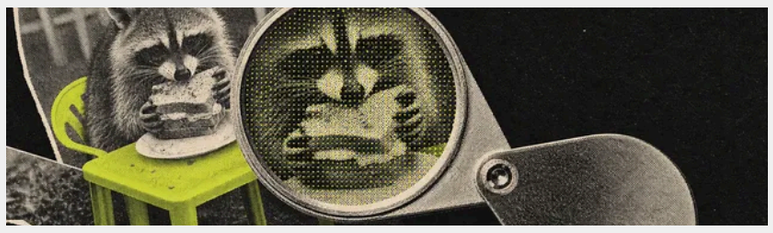
AI의 쓰임새는 1년 사이 크게 바뀌었다. 오픈AI 사례에 따르면 한때 코드 조각을 쓰거나 작은 버그를 찾던 도구가 이제 코드베이스와 서비스 전체를 혼자 만든다. 베이커 팀은 AI를 동료처럼 다룰 때 가장 잘 돌아간다고 본다. 연구자가 핵심 문제와 작업 환경, 결과의 품질과 의미를 평가할 방법을 정하면, 하나 또는 여러 에이전트에게 어디에 분석을 집중할지, 다음에 어떤 작업을 돌릴지 실시간으로 방향을 준다.

GPT-6 Astra에서 달라진 것은 개선을 쌓는 능력이다. 베이커에 따르면 에이전트는 의미 있는 변경을 찾는 데서 그치지 않고 그 성과들을 합치고 겹쳐 재귀적으로 개선한다. 오래 도는 작업 하나 안에서 시스템이 스스로 결과를 분석하고, 처음 제안서에서 합의한 기준에 비춰 판단하고, 사람이 매 회차를 들여다보지 않아도 방향을 바꾼다. 베이커는 "며칠 동안 돌아가야 하고, 여러 데이터 소스에서 끌어와야 하고, 무엇이 중요한지 미묘하게 판단해야 하고, 모든 것을 서로 엮어야 하는 작업을 정의할 수 있다"며 "그런 종합 분석이 이제 실제로 돌아간다"고 말했다.

AI

## 구글, SynthID Detector 일반 공개

**구글 딥마인드가 AI 생성 이미지·영상·음성의 워터마크를 확인하는 SynthID Detector를 전 세계에 영어로 일반 공개했다. 오픈AI와 엔비디아, 카카오 도구의 결과물까지 한 창구에서 검사하고 애플도 곧 합류한다.**



1 구글 딥마인드가 10월 7일 SynthID Detector를 대기자 명단 없이 전 세계에 영어로 일반 공개했다.

2 구글과 오픈AI, 엔비디아, 카카오 도구로 만든 이미지·영상·음성을 검사하고 애플 인텔리전스 결과물도 곧 지원한다.

3 SynthID 워터마크는 1800억 장이 넘는 이미지·영상과 24만 년 분량의 음성에 들어갔고, 검색·제미나이·크롬의 검증 기능은 하루 100만 건 넘는 요청을 처리한다.

구글이 AI로 만든 이미지와 영상, 음성을 누구나 가려 볼 수 있는 SynthID Detector를 일반 공개했다. 구글 딥마인드는 10월 7일 이 검사 도구를 이날부터 전 세계에 영어로 연다고 밝혔다. 이용자는 synthid.com에 파일을 올려 구글 AI나 파트너사 도구로 만든 것인지 확인할 수 있다. 파트너사에는 오픈AI와 엔비디아, 카카오가 들어 있고 애플도 곧 합류한다.

이 도구는 지난해 초기판으로 먼저 나왔다. 당시에는 언론인과 연구자 같은 미디어 전문가가 대기자 명단을 거쳐야 쓸 수 있었다. 보도에 따르면 이번 공개로 대기자 명단이 사라졌고, 구글이나 애플, 오픈AI 계정으로 로그인하면 바로 쓸 수 있다. 구글 딥마인드에서 과학·전략 이니셔티브를 맡은 푸시밋 콜리 부사장은 발표 글에서 "AI 도구가 더 좋아지고 쓰기 쉬워질수록, 온라인에서 마주치는 콘텐츠에 대한 분명한 맥락이 무엇을 믿을지 판단하는 데 도움이 된다"고 썼다.

바탕은 2023년에 나온 SynthID 워터마크다. 생성 시점에 이미지·영상의 픽셀이나 음성 신호 안에 사람 눈과 귀로는 알아챌 수 없는 신호를 심는다. 눈에 보이는 표식이나 메타데이터에 기대지 않기 때문에 잘라내기와 압축, 색 보정, 음성 처리를 거친 뒤에도 남도록 설계됐다고 보도됐다. 구글에 따르면 지금까지 1800억 장이 넘는 이미지·영상과 24만 년 분량의 음성에 이 워터마크가 들어갔다.

AI

## 에브리, 슬랙용 에브리 에이전트 공개

**미디어·소프트웨어 회사 에브리가 Claude Managed Agents 위에 만든 사내 공유 에이전트를 6개월 사내 사용 끝에 외부에 내놓았다. 슬랙 스레드에서 일하고 고친 내용을 스킬로 저장해 한 사람의 요령을 회사 전체가 쓰게 한다.**



1 에브리가 슬랙에서 @Every 태그로 일을 맡기는 공유 AI 동료 에브리 에이전트를 공개했고, 앤스로픽이 10월 6일 공식 X 영상으로 소개했다.

2 에이전트는 1000개가 넘는 도구 연결과 사전 탑재 스킬, 새 모델이 나오면 팀에 쓸모를 알려 주는 프런티어 알림을 갖췄고 토큰 비용에 마진을 붙이지 않는다.

3 에브리는 개인 에이전트를 나눠 준 플러스 원 실험 끝에 공유 에이전트 하나가 회사를 더 빨리 바꾼다고 판단했고, 인프라는 Claude Managed Agents에 맡겼다.

미국 미디어·소프트웨어 회사 에브리(Every)가 슬랙 안에서 회사 전체가 함께 쓰는 AI 동료 에브리 에이전트(Every Agent)를 공개했다. 앤스로픽의 에이전트 실행 기반인 Claude Managed Agents 위에 만든 이 에이전트는 에브리가 약 6개월 동안 사내에서 먼저 써 온 도구다. 앤스로픽은 10월 6일(미국 시간) 공식 X 계정에 3분 가까운 영상을 올려 "사내에서 자리를 잡자 구독자에게 내놓았다"고 소개했다. 이 게시물은 하루도 안 돼 조회 20만9000회를 넘겼다.

사용법은 단순하다. 어느 슬랙 채널에서든 @Every를 태그하고 할 일을 적으면 에이전트가 그 스레드 안에서 일을 처리한다. 결과가 틀리면 스레드에서 바로 고쳐 주고, 에이전트는 그 수정을 다음을 위해 저장할 수 있다. 구글 드라이브와 지메일, 노션, 피그마, 깃허브, 허브스팟을 포함해 1000개가 넘는 도구를 한 번 연결하면 팀원별로 어떤 연결을 쓸지 정할 수 있다. 에브리가 회사를 운영하며 쓰는 컴파운드 엔지니어링(Compound Engineering)과 컴파운드 라이팅(Compound Writing) 같은 스킬이 미리 깔려 있고, 반복할 만한 작업 흐름은 @Every에게 스킬로 저장하라고 하면 누구나 다시 돌릴 수 있다.

에브리가 내세운 차별점은 프런티어 알림(Frontier Alerts)이다. 새 모델이 나오면 에이전트가 그 팀이 지금 쓰는 도구와 작업 흐름을 훑어보고 그 모델이 우리 일에 쓸모가 있는지 알려 준다. 에브리는 에이전트가 쓰는 토큰에 마진을 붙이지 않는다고 밝혔다. 회사가 모델 회사에 내는 값을 이용자가 그대로 낸다는 뜻이다. 팀원 한 명마다 무료 크레딧이 주어지고, 설치에는 무료 에브리 계정과 슬랙 워크스페이스 소유자의 허가가 필요하다.