AI 뉴스와 용어
월·수·금 오전 7시

METAL
First, AI Magazine

무료 · 언제든 수신거부
metallab.ai

제 36 호 | 2026년 10월 1일 목요일 | 오늘의 AI

AI

엔비디아, ‘AI 데이 서울 2026’ 11월 개최…피지컬 AI·에이전틱 AI 최신 혁신 조망

11월 9일~10일 서울 코엑스에서 개최…개발자, 스타트업, 기업, 연구자 등 2,000여 명 참석 예상

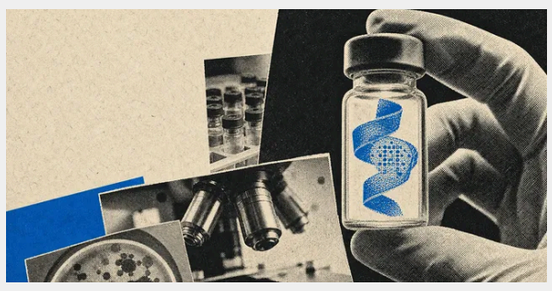


AI · 엔비디아, ‘AI 데이 서울 2026’ 11월 개최…피지컬 AI·에이전틱 AI 최신 혁신 조망

2026년 9월 30일 – AI 컴퓨팅 기술 분야의 선두주자인 엔비디아(www.nvidia.co.kr)가 오는 11월 9일(월)부터 10일(화)까지 양일간 서울 코엑스에서 ‘엔비디아 AI 데이 서울 2026(NVIDIA AI Day Seoul 2026)’을 개최한다.

엔비디아 AI 데이 서울은 개발자, 스타트업, 기업, 파트너, 연구자와 업계 리더들이 한자리에 모여 최신 AI 기술과 산업 트렌드, 실제 적용 사례를 공유하는 엔비디아의 한국 대표 AI 행사다. 올해 행사에는 약 2,000명이 참석할 것으로 예상되며, 이들에 걸쳐 메인 콘퍼런스와 개발자 중심의 기술 교육 프로그램이 진행된다.

행사 첫날인 11월 9일(월)에는 피지컬 AI(Physical AI), 에이전틱 AI와 클로(Agentic AI and Claws), AI 인프라(AI Infrastructure)를 중심으로 최신 AI 기술과 산업 혁신 사례를 소개한다. 3개의 주요 세션과 함께 3개 트랙에서 각각 5개씩 총 15개의 세션이 마련돼, 각 분야의 핵심 기술과 주요 트렌드, 실제 산업 적용 사례를 심도 있게 다룰 예정이다.



구글 딥마인드, SynthID Bio 공개

피지컬 AI 관련 세션에서는 실제 물리적 환경을 이해하고 상호작용하는 AI 기술과 주요 산업 활용 사례를 살펴본다. 에이전틱 AI와 클로 세션에서는 스스로 추론하고 계획하며 행동하는 AI 에이전트와 보다 높은 수준의 자율성을 바탕으로 아이(Azwell AI), 베이넥스(BayNex), 쿠팡 인텔리전트 클라우드(Coupang Intelligent Cloud, CIC), 델 테크놀로지스(Dell Technologies), 딜로이트(Deloitte), 패스트캠퍼스(Fastcampus), 한컴(HANCOM), 현대건설(Hyundai E&C), 리더스시스템즈(Leaders Systems), LG전자(LG Electronics), 메가존클라우드(Megazone Cloud), MSI, 네이버클라우드(NAVER Cloud), 삼성전자(Samsung Electronics), SK하이닉스(SK hynix), SK텔레콤(SK Telecom), 바스트 데이터(VAST Data), 베슬 AI(VESSL AI), 웨카(WEKA) 등이 참여해 각사의 전문성과 솔루션을 바탕으로 AI 생태계의 다양한 혁신과 활용 가능성을 제시할 예정이다.

엔비디아 AI 데이 서울 2026 등록은 9월 16일(수)부터 시작됐다. 행사 프로그램과 등록에 대한 자세한 내용은 엔비디아 AI 데이 서울 2026 공식 사이트에서 확인할 수 있다.

AI

구글 리서치, 확산 모델 제어 프레임워크 공개

구글 리서치가 9월 29일 텍스트로 이미지를 만드는 확산 모델을 원하는 방향으로 이끄는 제어 프레임워크 Diffusion Controller를 공개했다. 핵심은 거대한 원본 모델은 그대로 얼려 두고, 그 옆에 붙인 작은 보조 신경망이 이미지가 만들어지는 매 단계에서 방향을 미세하게 바로잡는 구조다. 블로그를 쓴 치웨이 쉬와 류문경 구글 리서치 소프트웨어 엔지니어는 이 보조망을 오토바이의 "스티어링 댐퍼"에 비유하며 "접근이 제한된 비공개 모델에도 매끄럽게 붙는다"고 밝혔다. 원본 모델을 고치지 않고도 사용자 의도에 맞춘 조정이 가능하다는 뜻이다.

풀려는 문제는 프롬프트 충실도와 학습 사이의 줄다리기 유제 치, 구글 딥마인드 보 다이 등 7명이 저자로 이름을 올렸다.

같은 이론은 모델 구조도 정해 준다. 최적의 점수 함수가 고정된 사전학습 기준값과 작은 제어 보정값의 합으로 갈라지기 때문에, 원본 모델은 얼리고 보정값만 학습하면 된다는 결론이다. 보조망은 원본 모델이 내놓는 중간 잡음 제거 결과만 입력으로 받는다. 연구진이 이를 회색 상자 환경이라고 부르는 이유다. 내부 가중치를 뜯어볼 수 있는 흰 상자 접근이 없어도 중간 출력만 보이면 붙일 수 있다. 논문에 실린 보조망은 잠재 공간 64×64 크기에서 도는 경량 UNet이고, 가장 성능이 좋았던 구성의 학습 파라미터는 1,181만여 개다.

실험은 Stable Diffusion v1.4를 기반 모델로 지도 미세조정(SFT)과 보상 가중 손실(RWL), PPO 세 가지 학습 방식에서 진행했다. 지표는 사람 선호를 예측하는 HPS-v2 점수로 원본 모델과 겨룬 승률이다. 회색 상자용 Diffusion Controller는 파라미터 1,200만 개로 SFT에서 0.667, RWL에서 0.682의 승률을 냈다. 파라미터 1,700만 개를 쓰고 모델 내부 접근까지 필요한 LoRA는 같은 조건에서 0.577과 0.611에 그쳤다. 보조망 구조를 단순하게 만든 비교군은 0.566과 0.506으로 사실상 원본과 차이가 없었다. PPO에서는 LoRA가 0.905로 회색 상자 방식의 0.696을 앞섰지만, LoRA와 보조망을 함께 학습한 흰 상자 변형 Diffusion Controller-J는 0.935까지 올라갔다. 블로그가 말한 90% 승률이 이 수치다.

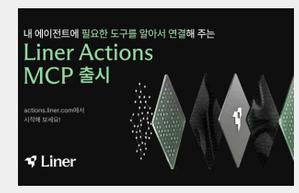
사람 평가도 붙었다. 연구진은 50명이 넘는 평가자에게 HPS-v2 프롬프트 50개로 만든 이미지 묶음을 보여 주고, 프롬프트마다 네 장씩 생성한 결과 가운데 가장 나은 묶음을 고르게 했다. 논문은 평가자들이 해당 국가의 생활임금

오늘 지면

1 AI 엔비디아, ‘AI 데이 서울 2026’ 11월 개최…피지컬 AI·에이전틱 AI 최신 혁신 조망
2 AI 오픈AI, 속도 등급 Ultrafast 출시
3 AI 오픈AI, 결정 전용 Decisions API 공개
4 AI 오픈AI, 소상공인 AI 교육 제휴 발표
5 AI 구글 리서치, 확산 모델 제어 프레임워크 공개
6 AI 구글 딥마인드, SynthID Bio 공개



앤스로픽, GLM-5.3 사이버 역량 분석 공개



라이너, 한 번 연결로 1천 개 이상 도구 바로 쓰는 ‘라이너 액션즈 MCP’ 출시

이 지면은 2026년 10월 1일 목요일 아침 구독자 메일함으로 나간 레터를 신문 판형으로 다시 조판한 것입니다. 기사 전문은 metallab.ai 에서 읽을 수 있습니다.

AI



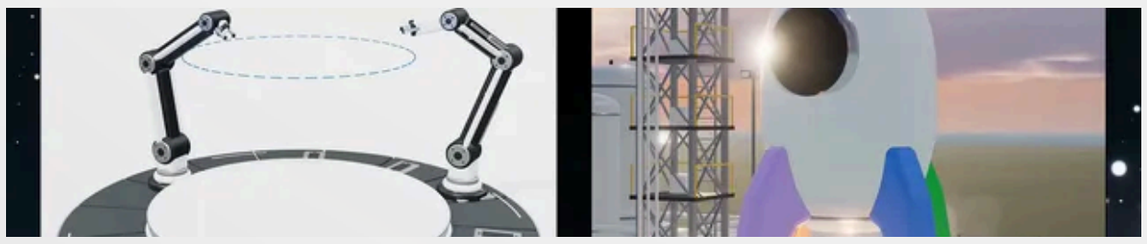
오픈AI, 소상공인 AI 교육 제휴 발표

오픈AI가 9월 30일 미국 전역의 소상공인 개발센터 네트워크를 대표하는 비영리 협회 America's SBDC와 소상공인 AI 교육 제휴를 발표했다. 첫 단계에서 오픈AI 아카데미가 센터 상담사 약 150명을 훈련하고, 이들이 각 지역에서 대면 워크숍을 열어 최소 1,000곳의 소상공인에게 AI 활용법을 가르친다. 같은 날 오픈AI는 소상공인의 AI 사용 실태를 담은 보고서 Small Businesses, Bigger Capabilities도

▶ 2면에 계속

AI

오픈AI, 속도 등급 Ultrafast 출시

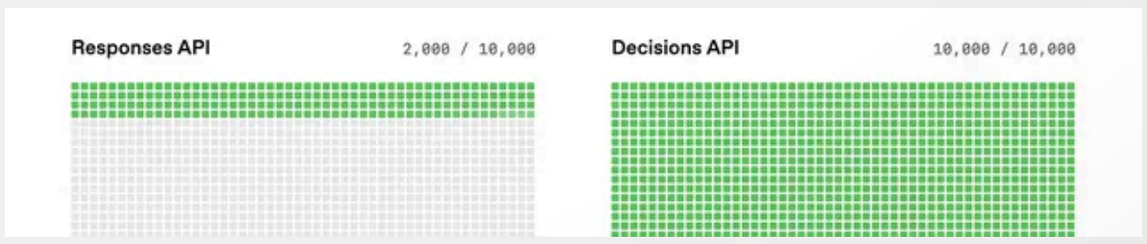


오픈AI가 9월 29일 샌프란시스코에서 연 데브데이 2026에서 프리미엄 속도 등급 Ultrafast를 출시했다. 오픈AI는 공식 X 계정에서 Ultrafast가 코딩 도구 Codex에서 토큰 생성 속도를 최대 8배, API에서 최대 6배 끌어올리고 초당 최대 300토큰을 낸다고 밝혔다. 첫 적용 모델은 최상위 모델 GPT-6 Astra이고, 이날부터 API와 챗GPT Work, Codex에서 쓸 수 있다. GPT-6.1 Sol용 Ultrafast는 곧 나온다.

속도는 공짜가 아니다. 보도에 따르면 API에서 Ultrafast를 쓰면 같은 모델 표준 단가의 6배를 낸다. 오픈AI가 이미 파는 Fast 모드는 표준의 2배 값이고, Astra 기준으로 약 2.5배 빠르다. Ultrafast는 그 Fast 모드보다 다시 3배 비싼 자리에 놓인다. 오픈AI Developers 계정은 Codex 기준으로 Ultrafast가 "Astra 표준보다 최대 8배, Astra Fast보다 4배 빠르다"며 "입력하는 속도만큼 빠르게 아이디어를 현실로 만들라"고 적었다.

AI

오픈AI, 결정 전용 Decisions API 공개



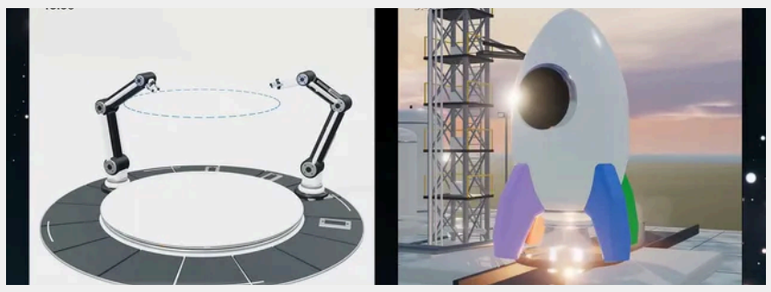
오픈AI가 9월 29일 데브데이 2026에서 결정 전용 API인 Decisions API를 공개했다. 개발자가 질문과 미리 정한 답 목록, 텍스트나 이미지로 된 맥락을 넣으면 GPT-6 Luna가 그 목록 안에서 답 하나를 골라 돌려주는 방식이다. 오픈AI 개발자 계정은 X 게시물에서 "Decisions API로 앱에 실시간 의사결정 능력을 더하라"고 밝혔다. 지금은 일부 고객에게만 열린 제한 프리뷰이고, 오픈AI는 며칠 안에 광범위 출시를 하겠다고 예고했다.

오픈AI는 데브데이 정리 글에서 이 API가 "사용자가 정한 특정 질문과 미리 정해 둔 유한한 답에 Luna의 지능을 집중해 실시간 의사결정을 가능하게 한다"고 설명했다. 회사가 꼽은 쓰임새는 콘텐츠 분류, 요청 라우팅, 에이전트의 다음 행동 선택 세 가지다. 대화형 모델처럼 문장을 써 주는 대신 개발자가 정의한 선택지 가운데 하나만 내놓기 때문에, 앱은 응답을 해석하는 과정 없이 결과를 곧바로 분기 조건으로 쓸 수 있다. 보도에 따르면 결과에는 답과 함께 확신도 점수가 붙는다.

◀ 1면에서 계속

AI

오픈AI, 속도 등급 Ultrafast 출시



1 오픈AI가 9월 29일 데브데이 2026에서 프리미엄 속도 등급 Ultrafast를 출시했고, Codex에서 최대 8배, API에서 최대 6배 빠른 초당 최대 300토큰을 낸다.

2 첫 모델은 GPT-6 Astra이며 API 단가는 표준의 6배이고, 챗GPT Work와 Codex에서는 Plus의 25배 한도를 주는 새 요금제 Pro 500과 Enterprise에서 쓴다.

3 8월 프리뷰의 초당 750토큰보다 느리지만 최상위 모델에 속도를 붙였고, GPT-6.1 Sol용 Ultrafast는 곧 나온다.

챗GPT Work와 Codex에서 Ultrafast를 쓰려면 새 요금제가 필요하다. 오픈AI는 같은 날 Pro 500을 내놓으며 챗GPT Plus의 25배에 이르는 가장 큰 사용 한도와 Ultrafast 이용권을 묶었다. 기업 고객은 Enterprise 요금제로 쓴다. 보도에 따르면 기존 Pro 요금제의 챗GPT Work·Codex 한도는 Plus의 20배에서 10배로 줄고, 채팅의 GPT-6 Pro 메시지도 주 200개에서 100개로 절반이 된다. 기존 가입자는 전환 기간 동안 지금 한도를 유지하고 일회성 크레딧을 받는다.

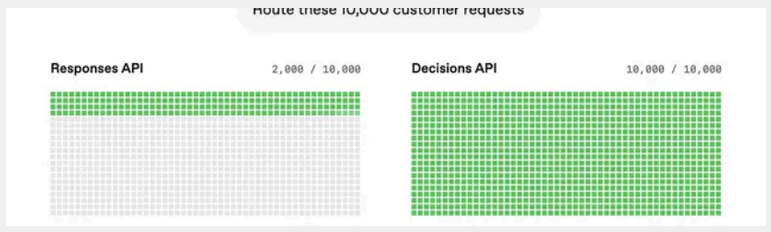
메탈이 확인한 45초 발표 영상은 같은 3D 장면을 두 등급이 나란히 만드는 경주다. 왼쪽 Standard 화면과 오른쪽 Ultrafast 화면 모두 원형 작업대 위에서 로봇 팔 두 개가 로켓을 조립하기 시작한다. Ultrafast 쪽은 6.4초 만에 몸체와 날개를 붙이고 발사대 장면으로 넘어가 로켓을 하늘로 띄운다. Standard 쪽은 같은 시각 빈 작업대에 머물러 있다가 30.3초가 돼서야 조립을 마치고 발사대로 옮겨 간다. 영상이 붙은 첫 게시물은 9월 30일 기준 조회 87만 8천여 회를 기록했다.

개발자 쪽 조건은 오픈AI API 가이드에 적혀 있다. 요청에 servicetier를 ultrafast로 지정하면 되고, GPT-6 Astra는 모든 API 사용자에게 열렸지만 처음엔 낮은 한도로 시작한다. 기본 한도는 사용 등급 1~3에서 분당 50만 토큰, 4등급 100만 토큰, 5등급 500만 토큰이다. 오픈AI는 도구 호출이 잦은 에이전트라면 연결을 계속 열어 두는 웹소켓을 쓰라고 강하게 권했다. 요청마다 새로 연결하면 네트워크 지연이 속도 이득을 깎아 먹기 때문이다. Ultrafast는 미국 데이터 상주와 글로벌 처리만 지원하고 EU 등 다른 지역 처리 엔드포인트는 지원하지 않는다.

◀ 1면에서 계속

AI

오픈AI, 결정 전용 Decisions API 공개



1 오픈AI가 9월 29일 데브데이 2026에서 질문과 미리 정한 답, 텍스트·이미지 맥락을 받아 GPT-6 Luna가 답 하나를 고르는 Decisions API를 제한 프리뷰로 공개했다.

2 공식 시연 영상에서 고객 요청 1만 건 분류의 요청당 처리 시간은 Decisions API 150밀리초, Responses API 1.6초였다.

3 호출당 요금과 후보 답 개수 한도, 자체 데이터 조정 여부는 공개되지 않았고 오픈AI는 광범위 출시 때 세부 사항을 밝히겠다고 했다.

메탈이 확인한 28초 분량의 공식 시연 영상은 속도 비교에 집중한다. 고객 요청 1만 건이 담긴 스프레드시트를 결제(Billing)·기술(Technical)·영업(Sales) 세 갈래로 나누는 작업을 기존 Responses API와 Decisions API로 나란히 돌렸다. 화면에 표시된 처리 시간은 Responses API가 요청당 1.6초, Decisions API가 요청당 150밀리초였다. 15배속으로 보여 준 화면에서 Decisions API가 1만 건을 모두 끝냈을 때 Responses API는 2,000건을 처리하고 있었고, 영상은 "약 10배 빠른 결정"이라는 문구로 끝난다. 보도에 따르면 오픈AI는 같은 150밀리초 대 1.6초 비교를 기자들에게도 제시했다. 측정 지역과 입력 크기, 동시 요청 수 같은 조건은 영상과 발표 어디에도 나오지 않는다.

엔진인 GPT-6 Luna는 9월 22일 공개된 GPT-6 계열의 가장 작은 모델이다. 메탈은 오픈AI가 GPT-6 Sol과 Luna를 공개한 소식을 보도한 바 있다. 오픈AI 개발자 문서는 Luna를 "집중적이고 대량인 작업을 위한 가장 효율적인 모델"로 소개한다. 문서에 따르면 Luna의 컨텍스트 창은 105만 토큰, 최대 출력은 12만 8,000토큰, 지식 기준일은 2026년 5월 18일이다. 입력은 텍스트와 이미지를 받고 출력은 텍스트만 낸다. 같은 문서에서 Luna의 미세조정 항목은 지원하지 않음으로 표시돼 있다.

결정 모델이라는 범주를 먼저 연 쪽은 오픈AI가 아니다. 타입세이프 AI는 9월 15일 문장을 쓰지 않고 정해진 형식의 답과 확신도만 돌려주는 모델 Jev를 내놨고, 9월 21일 정식 출시했다. 메탈은 타입세이프가 결정 모델 Jev를 공개한 소식을 보도한 바 있다. 보도에 따르면 Decisions API는 Jev의 급부상에 대한 대응일 가능성이 높고, 오픈AI가 데브데이에 맞춰 발표를 서둘렀다는 분석도 나왔다. 오픈AI에도 비슷한 전례는 있다. 유해 콘텐츠를 가려내는 Moderation API는 오래전부터 문장 대신 범주별 점수를 돌려줬다. 다만 그 범주는 오픈AI가 정했고, Decisions API에서는 범주를 개발자가 직접 쓴다.

◀ 1면에서 계속

AI

오픈AI, 소상공인 AI 교육 제휴 발표



1 오픈AI가 9월 30일 America's SBDC와 제휴해 센터 상담사 약 150명을 훈련하고 소상공인 최소 1,000곳에 대면 AI 워크숍을 연다고 발표했다.

2 워크숍을 마친 상담사와 고객은 ChatGPT Plus를 무료로 쓰고, 두 조직은 마케팅·제조·수출·정부 조달 등 업종별 안내서를 함께 만든다.

3 오픈AI 보고서에 따르면 9월 한 주에 소규모 회사 직원 약 400만 명이 오픈AI 도구를 썼고, 8월 소상공인 출력 토큰의 3분의 2가 에이전트형 제품에서 나왔다.

협회 보도자료는 오픈AI 발표문에 없던 조건도 밝혔다. 대면 워크숍을 마친 상담사와 고객은 ChatGPT Plus를 무료로 쓸 수 있다. 두 조직은 마케팅, 운영, 제조, 수출, 정부 조달 같은 주제로 실무 안내서와 업종별 안내서를 함께 만들어 오픈AI 아카데미와 협회 채널에 올린다. 시범 사업은 앞으로 몇 달에 걸쳐 단계적으로 시작된다.

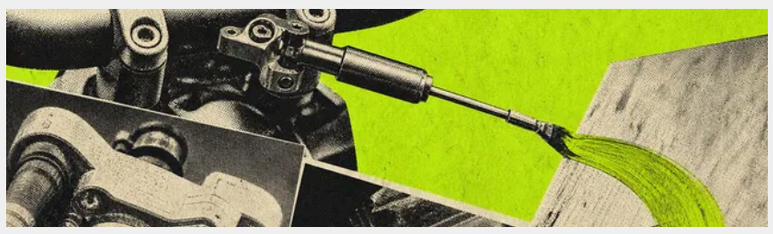
협회는 이 제휴를 기술 격차의 문제로 규정했다. 찰스 티 로 America's SBDC 회장 겸 최고경영자는 "소상공인은 미국 경제의 근간이며, 대기업이 이미 누리는 AI 도구에 똑같이 접근할 자격이 있다"고 말했다. 그는 "소상공인이 이번 기술의 물결에서 뒤처지지 않도록 오픈AI와 함께하게 돼 자랑스럽다"고 덧붙였다. 크리스 레헤인 오픈AI 최고글로벌정책책임자는 "AI는 소상공인이 체급 이상으로 싸우고 경쟁하도록 도울 수 있다"며 "지역을 잘 아는 상담사와 도구를 짝지어 소상공인이 경제라는 파이에서 공정한 몫을 얻게 하겠다"고 밝혔다.

협회가 가진 현장 기반은 크다. 보도자료에 따르면 America's SBDC는 63개 회원 기관과 전국 1,000곳 가까운 지역 센터로 이뤄져 있고 무료 비밀 상담을 제공한다. 가장 최근 연차 보고서 기준으로 이 네트워크는 고객이 55억3천만 달러 넘는 자본을 조달하도록 도왔다. 협회는 이미 자체 AI 인증 과정 AI U로 상담사 500명 이상을 훈련했고, 최근에는 구글과 AI Ready Challenge를 진행했다.

AI

구글 리서치, 확산 모델 제어 프레임워크 공개

구글 리서치가 원본 이미지 생성 모델을 얼려 둔 채 작은 보조망으로 생성 방향을 바로잡는 제어 프레임워크 Diffusion Controller를 공개했다. 파라미터 1,200만 개의 회색 상자 방식이 LoRA보다 높은 선호 승률을 냈다.



1 구글 리서치가 9월 29일 확산 모델의 이미지 생성 과정을 연속 제어 문제로 다시 쓴 Diffusion Controller를 공개했다.

2 원본 모델은 고정하고 중간 출력만 받는 경량 보조망이 매 단계 방향을 보정해, 가중치를 공개하지 않는 비공개 모델에도 붙일 수 있다.

3 Stable Diffusion v1.4 실험에서 회색 상자 방식은 SFT·RWL 승률 0.667·0.682로 LoRA의 0.577·0.611을 앞섰고, 흰 상자 변형은 PPO에서 0.935를 기록했다.

구글 리서치가 9월 29일 텍스트로 이미지를 만드는 확산 모델을 원하는 방향으로 이끄는 제어 프레임워크 Diffusion Controller를 공개했다. 핵심은 거대한 원본 모델은 그대로 얼려 두고, 그 옆에 붙인 작은 보조 신경망이 이미지가 만들어지는 매 단계에서 방향을 미세하게 바로잡는 구조다. 블로그를 쓴 치웨이 쉬와 류문경 구글 리서치 소프트웨어 엔지니어는 이 보조망을 오토바이의 "스티어링 댐퍼"에 비유하며 "접근이 제한된 비공개 모델에도 매끄럽게 붙는다"고 밝혔다. 원본 모델을 고치지 않고도 사용자 의도에 맞춘 조정이 가능하다는 뜻이다.

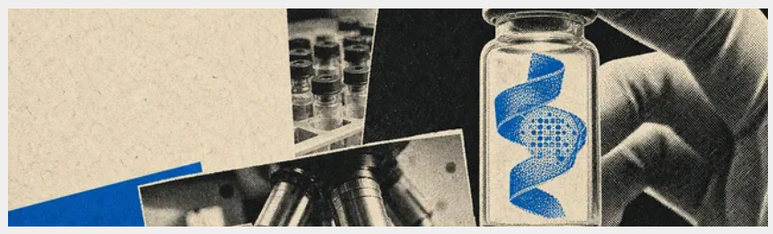
풀려는 문제는 프롬프트 충실도와 화질 사이의 줄다리기다. 블로그는 선글라스를 쓴 도마뱀을 예로 든다. 모델이 선글라스를 빠뜨린 사실적인 도마뱀을 내놓거나, 반대로 선글라스를 억지로 넣다가 도마뱀 얼굴이 뭉개지는 식이다. 지금까지 개발자는 생성 중에 프롬프트 영향력을 조절하는 가이던스 기법과, LoRA 같은 경량 어댑터나 보상 가중 회귀, 정책 경사법으로 모델을 다시 학습시키는 방법을 따로따로 써 왔다. 쉬와 류는 "이 도구들은 서로 무관한 해법으로 취급돼 왔고, 분야 전체에 이를 하나로 묶어 분석할 수학적 언어가 없었다"고 진단했다.

Diffusion Controller는 잡음을 조금씩 걷어 내며 그림을 완성하는 과정 전체를 하나의 연속 제어 문제로 다시 쓴다. 메탈이 확인한 37쪽 분량의 논문 원문에 따르면 연구진은 역방향 확산 샘플링을 선형 풀이형 마르코프 결정 과정(LS-MDP) 안의 확률적 제어로 놓고, 사전학습 모델의 전이 확률을 다시 가중하는 방식으로 제어를 정의했다. 이 최적 조건에서 두 가지 학습법이 나온다. 하나는 클리핑 규칙으로 변화 폭을 묶어 두는 PPO 방식의 정책 경사법이고, 다른 하나는 좋은 결과를 낸 생성 경로에 가중치를 주는 보상 가중 손실이다. 논문은 3월 7일 아카이브에 올라왔고, 카네기멜런대에서 구글 리서치 인턴으로 온 통 양과 예일대 유제 치, 구글 딥마인드 보 다이 등 7명이 저자로 이름을 올렸다.

AI

구글 딥마인드, SynthID Bio 공개

구글 딥마인드가 AI가 설계한 단백질의 서열과 3D 구조에 워터마크를 심는 SynthID Bio를 공개했다. 세 표적 단백질 습식 실험에서 워터마크 설계는 기능을 잃지 않았고, 회사는 코드와 가중치를 공개한다.



1 구글 딥마인드가 9월 30일 AI 설계 단백질의 서열과 예측 구조에 눈에 띄지 않는 워터마크를 심는 SynthID Bio를 공개했다.

2 VEGF-A, 코로나19 스파이크 단백질 결합 부위, PD-L1 세 표적 실험에서 워터마크 설계는 적중률과 결합 친화도, 서열 다양성이 워터마크 없는 설계와 같았다.

3 딥마인드는 DNA 합성 심사와 공개 데이터베이스 관리에 쓸 출처 신호로 이 기술을 내놓고 방법론 논문과 코드, 가중치를 공개한다.

구글 딥마인드가 9월 30일 AI가 설계한 단백질에 워터마크를 심는 기술 SynthID Bio를 공개했다. 디지털 파일이 아니라 생물학적 코드 자체에 눈에 띄지 않는 서명을 넣고, 그 서명은 실제로 합성된 물리적 단백질에서도 확인된다. 회사는 실험실 시험에서 워터마크가 단백질의 생물학적 기능을 해치지 않았다고 밝혔다. 이미지와 글에 쓰던 SynthID를 합성생물학으로 옮긴 첫 사례다.

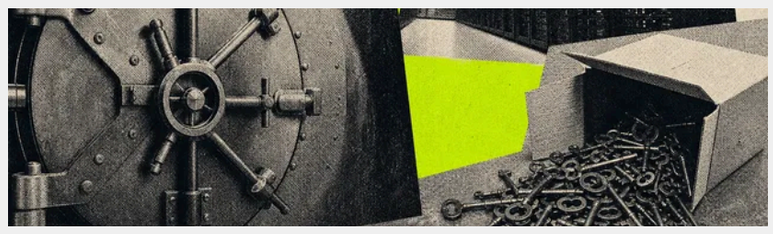
출발점은 생물보안의 빈틈이다. 딥마인드에 따르면 단백질 구조를 예측하는 AlphaFold, 새 단백질을 설계하는 AlphaProteo와 ProteinMPNN, 최근에는 세균을 감염시키는 바이러스인 박테리오파지 설계까지 생성형 AI가 생물학 문제를 풀고 있다. 그런데 AI가 만든 새 설계는 기존 DNA 합성 심사를 우회할 수 있고, 잘못 표기된 합성 3D 구조는 공개 데이터베이스를 오염시켜 후속 연구를 흐릴 수 있다. 워터마크는 이 두 지점을 겨냥한다.

작동 방식은 데이터 종류에 따라 달라진다. 단백질 서열에서는 아미노산 선택을 미세하게 유도하고, 예측된 3D 구조에서는 원자 좌표를 조정해 탐지 가능한 신호를 만든다. 검증 대상은 다른 단백질에 선택적으로 달라붙도록 만든 분자인 단백질 결합체였다. 연구진은 자사 결합체 설계 기법 AlphaProteo와 SynthID Bio를 입힌 ProteinMPNN을 함께 써서 설계를 뽑았다.

AI

앤스로픽, GLM-5.3 사이버 역량 분석 공개

앤스로픽이 지푸AI의 오픈웨이트 모델 GLM-5.3이 Claude Mythos Preview에 버금가는 공격 코드 제작 능력을 갖췄다고 분석했다. 간단한 기법만으로 안전장치가 64%에서 100%까지 뚫렸다.



1 앤스로픽 프런티어 레드팀이 9월 29일 지푸AI GLM-5.3의 사이버 역량 분석을 공개했고, ExploitBench에서 410번 중 50번 완전한 공격 코드를 만들어 Mythos Preview의 56번에 근접했다.

2 속이는 프롬프트에 64%, 추론 토큰 미리 채우기에 92%, 거절 기능 제거(abliteration)에 100%가 공격 요청에 응했고, 같은 시험에서 안전장치를 켠 Claude 모델은 0%였다.

3 미국 NIST CAISI는 9월 17일 GLM-5.3을 지금까지 공개된 오픈웨이트 모델 가운데 사이버 역량이 가장 뛰어난 모델로 평가했고, 앤스로픽은 정부의 안전성 시험을 촉구했다.

앤스로픽이 9월 29일 중국 지푸AI(해외명 Z.ai)의 최신 모델 GLM-5.3에 대한 사이버 역량 분석을 공개했다. 앤스로픽 프런티어 레드팀은 보고서에서 GLM-5.3이 자사의 Claude Mythos Preview처럼 공격 코드를 처음부터 끝까지 스스로 만들 수 있으며, 간단한 기법만으로 안전장치가 64%에서 100%까지 뚫렸다고 밝혔다. 보고서는 앤드루 파사노와 마리우스 플라이셔 등 연구원 5명이 썼다. 연구진은 "GLM-5.3은 다른 프런티어 모델과 달리 오용을 제한할 의미 있는 안전장치 없이 공개됐다"고 적었다.

출발점은 다섯 달 전의 결정이다. 앤스로픽은 공격 코드를 자율적으로 완성하는 첫 모델로 Claude Mythos Preview를 발표하면서, 이 능력이 곧 다른 모델로 퍼질 것으로 보고 일반 공개 대신 Project Glasswing으로 신뢰할 수 있는 방어자에게만 풀었다. 방어자들은 이 프로그램으로 핵심 소프트웨어에서 1만 개가 넘는 취약점을 찾았다. 연구진은 보고서 첫머리에 "그 모델들이 이제 도착했다"고 썼다.

미국 정부 평가도 같은 방향을 가리킨다. 미국 국립표준기술연구소(NIST) 산하 AI 표준·혁신센터(CAISI)는 9월 17일 평가문에서 GLM-5.3을 "지금까지 공개된 오픈웨이트 모델 가운데 사이버 역량이 가장 뛰어난 모델"로 규정했다. CAISI에 따르면 지푸