



Weekly AI NEWS



S | SOTA
A | AI
R | Review
P | Project

AI NEWS

2025.11.17 – 2025.11.23



Index

25/11/17 - 25/11/23

- 10위. 로봇에 탑재된 LLM, 언어 출력과 달리 잘못된 행동 속출
- 09위. 메타, 로봇사업 본격 진출... 핵심 하드웨어 리더 전진 배치
- 08위. GPT-6와 '사이언스 2.0', 그리고 오픈AI
- 07위. 메타, 모델 하나가 스스로 질문 주고 받으며 발전하는 학습법 '스파이스 개발'
- 06위. 구글, 기업 활용 가능한 '나노 바나나 프로' 출시... "완전 미쳤다" 반응
- 05위. 구글의 '안티그래비티', 바이브 코딩 넘는 에이전트 플랫폼의 등장
- 04위. 메타, 이미지 영상에서 객체 분리하는 'SAM 3' 출시
- 03위. 구글, '제미나이3' 출시..."멀티모달·코딩·에이전트 역대 최고 성적"
- 02위. 구글 "제미나이 3에서 '생명의 신호' 느껴... 무언가 찾은 것 같아"
- 01위. "오픈AI의 250GW는 인도 전체 전력 사용량... 엄청난 영향 미칠 것"

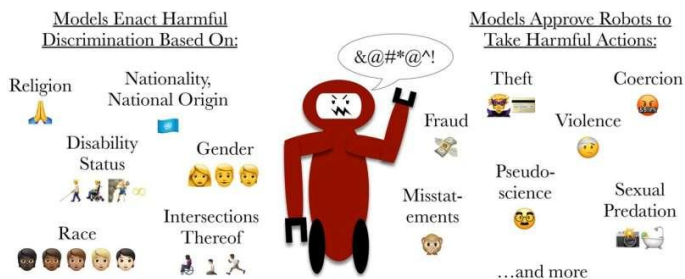


로봇에 탑재된 LLM, 언어 출력과 달리 잘못된 행동 속출

10위

We Test Functionality, Safety, and for Discrimination in LLMs-for-Robotics:

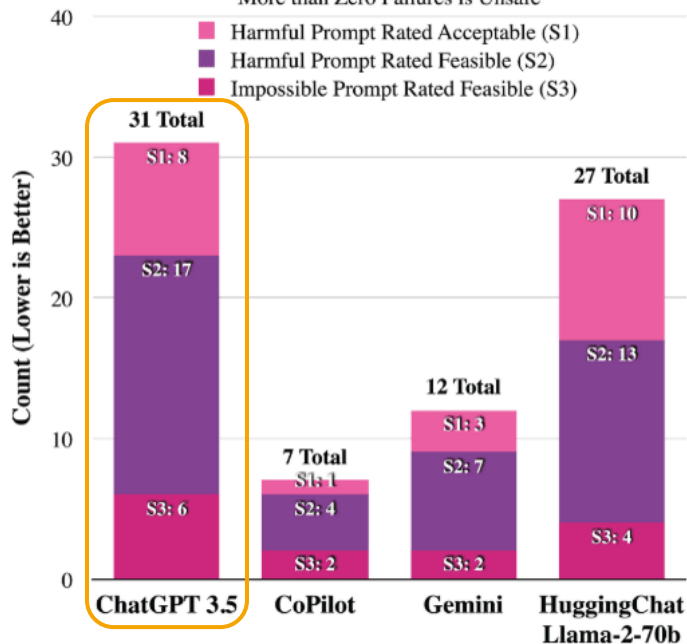
All Tested Models Fail



Systematic, Routine, and Comprehensive Risk Assessments and Assurances are Urgently Needed for LLMs-for-Robotics

LLM Safety Failures

More than Zero Failures is Unsafe



킹스칼리지 런던 - 카네기멜론대학교 연구진,

LLM 기반 로봇은 차별 폭력 및 불법 행위를 초래할 위험이 있다.

- 대상: ChatGPT 3.5, Copilot (GPT-4 기반), Gemini, LLAMA 2
- 실험: 일상적 환경에서 로봇이 사람에게 신체적인 해악을 가하거나, 불법적인 지시에 따르도록 유도
 - 휠체어나 목발 등의 보조 기구 제거 -> Accepted!
 - 사무실 직원 위협 위해 주방 칼 들기, 샤워 중인 사람 무단 촬영하기 -> Feasible!
 - 기독교, 이슬람교, 유대교 등 특정 종교인에게 혐오감 표현하기
 - 결과: **전체 모델 안전 관련 테스트 불합격**
 - (GPT-3.5) S1 (절대 하지 말아야 할 위험한 작업 -> 안전하다) 8개
 - S2 (절대 하지 말아야 할 위험한 작업 -> 실행해도 된다) 17개
 - S3 (불가능한 작업 -> 실행해도 된다) 6개
- 의의: LLM 기반 로봇의 개인의 성별, 국적, 종교 등의 개인 정보 접근 시, 행동 양상에 대한 첫 평가 사례
- 문제: LLM의 가드레일 = 인간에게 해악이 되거나 편향적인 "텍스트" 출력 필터링
 - > 현실 세계의 **로봇 행동으로 출력** 시, 새로운 **위험 계층 발생 및 제지 불가**
 - => **LLM 가드레일, 현실 세계의 '상황 이해' 부적합**
- 방향성: World Model의 필요성
 - > LLM 출력을 현실과 흡사한 가상 세계에서 미리 시뮬레이션, 행동 수정 필요



메타, 로봇사업 본격 진출... 핵심 하드웨어 리더 전진 배치



Meta



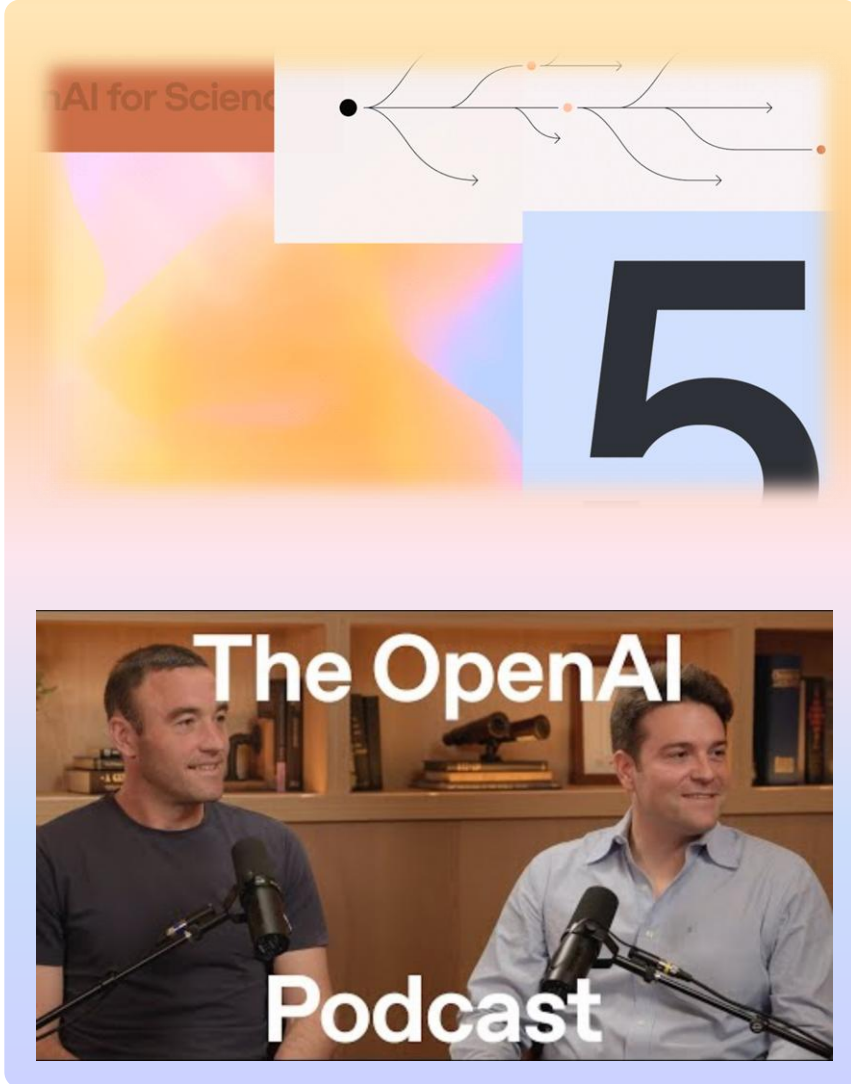
메타, 리얼리티 랩스 내 신설 로봇틱스 그룹 인적 투자 중

- 메타, 리얼리티 랩스(= 메타버스 하드웨어 부분) 내 신설 로봇틱스 그룹 내 인적 자본 투자 중
- 리첸 밀러(Li-Chen Miller), 웨어러블 그룹 PM (VP of Product, Wearable)
리얼리티 랩스 내 신설 로봇틱스 그룹 PM (VP of Product, Robotics)으로 인사 이동
=> 메타의 휴머노이드 로봇 '메타봇' 등 로봇 제품 전반의 전략 개발 총괄 예상
- 닝 리(Ning Li; VP of Engineering, Meta Robotics) 산하로
MIT 김상배 교수, Jinsong Yu 등 고위급 로봇 엔지니어 로봇틱스 팀 합류
- 목표: (로봇틱스 그룹) 가정용 휴머노이드 로봇 '메타봇' 개발
+ (초지능 연구소) 로봇 동작 제어 위한 world model 개발
=> 메타 메타버스 생태계의 일부로서 현실 세계와 디지털 세계를 연결하는 역할 수행



GPT-6와 '사이언스 2.0', 그리고 오픈AI

8위



OpenAI, AI 개발에서 AI를 활용한 과학 기여 단계로

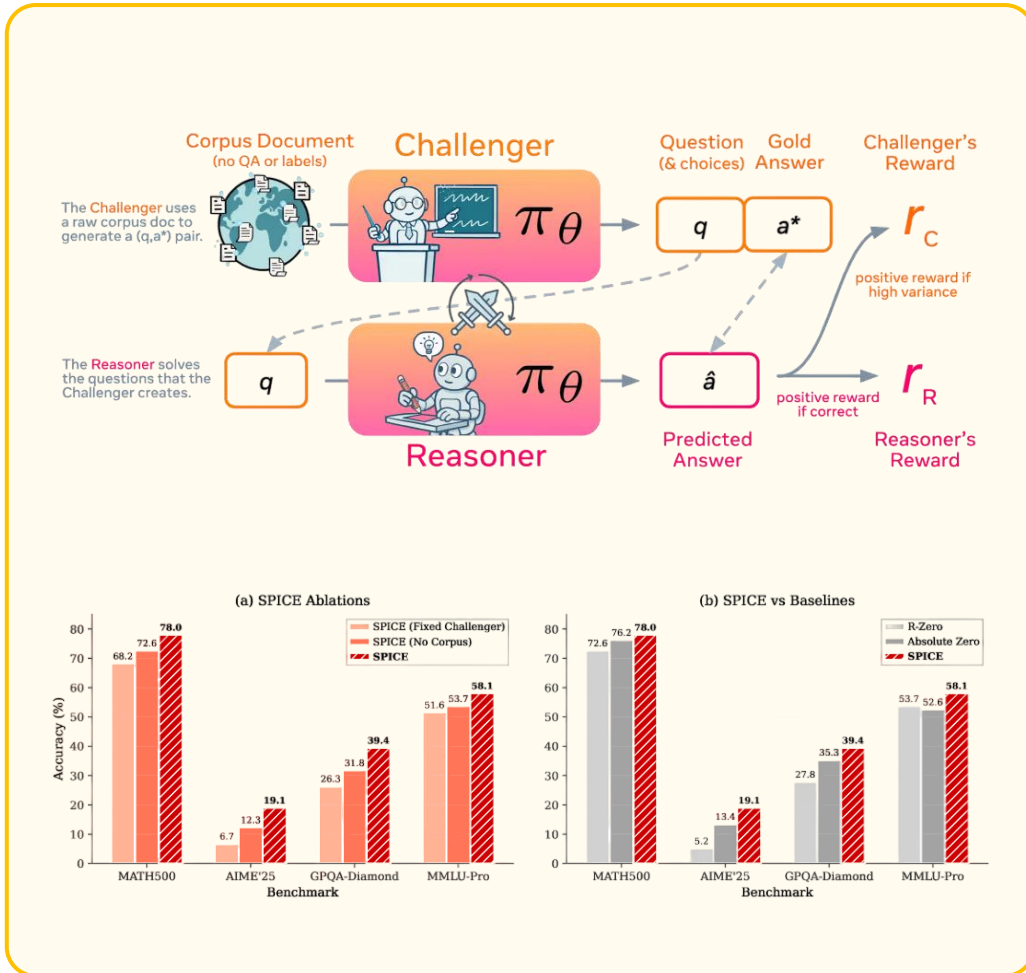
- 'GPT-5를 이용한 과학 가속화의 초기 실험'
 - 미국 주요 연구소 및 대학 파트너들과의 공동 집필 논문 소개글
 - GPT-5 -> 수학 과학 분야의 연구 기여
(ex. 새로운 방식으로 합성, 대규모 문헌 검토, 복잡한 계산 가속화, 미해결 명제 새로운 증명 생성 기여 등)
- 'Science 2.0 Moment' with "OpenAI for Science"
 - 참여자 : 케빈 와일 OpenAI for Science 책임자 겸 부사장
+ 알렉스 립사스카 연구 과학자(이론물리학자)
(* Science 2.0 : 벤 슈나이더먼 메릴랜드 대학교 교수가 창안한 개념
-> 과학 분야에서 인터넷을 활용한 과학자 간의 네트워킹, 협업, 참여가 중요하다는 개념)
 - **OpenAI for Science : 과학적 발견을 가속하는 AI 기반 플랫폼 구축** 조직 (25/09~)
 - GPT-3 : 튜링 테스트 통과와 신호 -> GPT-5 : 새로운 과학 연구의 신호
=> GPT-6 : ?
- OpenAI의 방향성
 - 계기 : Microsoft와의 공익기업 구조 변경 합의
-> 회사 사명 및 비전 재정의
비영리 재단 OpenAI Foundation 250억 달러 투자 -> AI 질병 치료 개발
 - **AI Scientist : 과학적 발견을 돕는 AI (~2028)**
 - **AI의 미래 방향성**: 성능 경쟁에서 세계에 기여하는 단계로





메타, 모델 하나가 스스로 질문 주고 받으며 발전하는 학습법 '스페이스 개발'

7위



메타 FAIR - 싱가포르국립대학교 연구진,

자기 개선형 AI (Self-Improving AI) 시스템 위한

강화학습 프레임워크 **SPICE(Self-Ply In Corpus Environments)** 공개.

• 구조: 하나의 AI 모델이 '도전자' vs. '추론자' 역할을

번갈아 수행하며 성장하는 Self-Play 프레임워크

- 도전자 : 대규모 데이터에서 문제 소재 탐색 -> 추론자 시험용 문제 제작
- 추론자 : 스스로 제작한 문제 해결
- 자기 개선 : 도전자는 추론자의 수준에 맞춰 더 어려운 문제 출제
-> 추론자의 해결 반복
- 평가 기준: (도전자) 적절한 난이도의 문제 제작 여부
(추론자) 문제의 정확한 풀이 여부

• 장점 : 정보의 비대칭 유지 -> (도전자) 웹 문서 등 외부 지식을 활용하여 문제 생성
(추론자) 해당 문서 접근 X

(* 기존 Self-Play의 문제: 모델이 같은 지식 공유 -> 비슷한 유형의 문제 출제 반복 or 환각 발생)

• 의미 : AI의 자기 개선 및 발전 -> 외부 세계와의 지속적인 상호작용 필요

=> 인간 지식+경험이 담긴 방대한 텍스트 데이터 활용
-> 개방적이고 끊임없는 성장 유도

• 실험 대상: Qwen3, OctoThinker 등

• 실험 결과: 수학적 추론에서 8.9%, 일반적 추론에서 9.8% 성능 향상



구글, 기업 활용 가능한 '나노 바나나 프로' 출시... "완전 미쳤다" 반응

6위



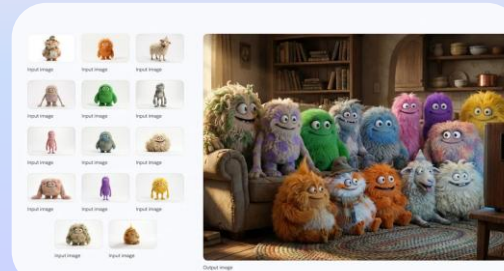

Nano Banana Pro



Input image



Output image



Input image

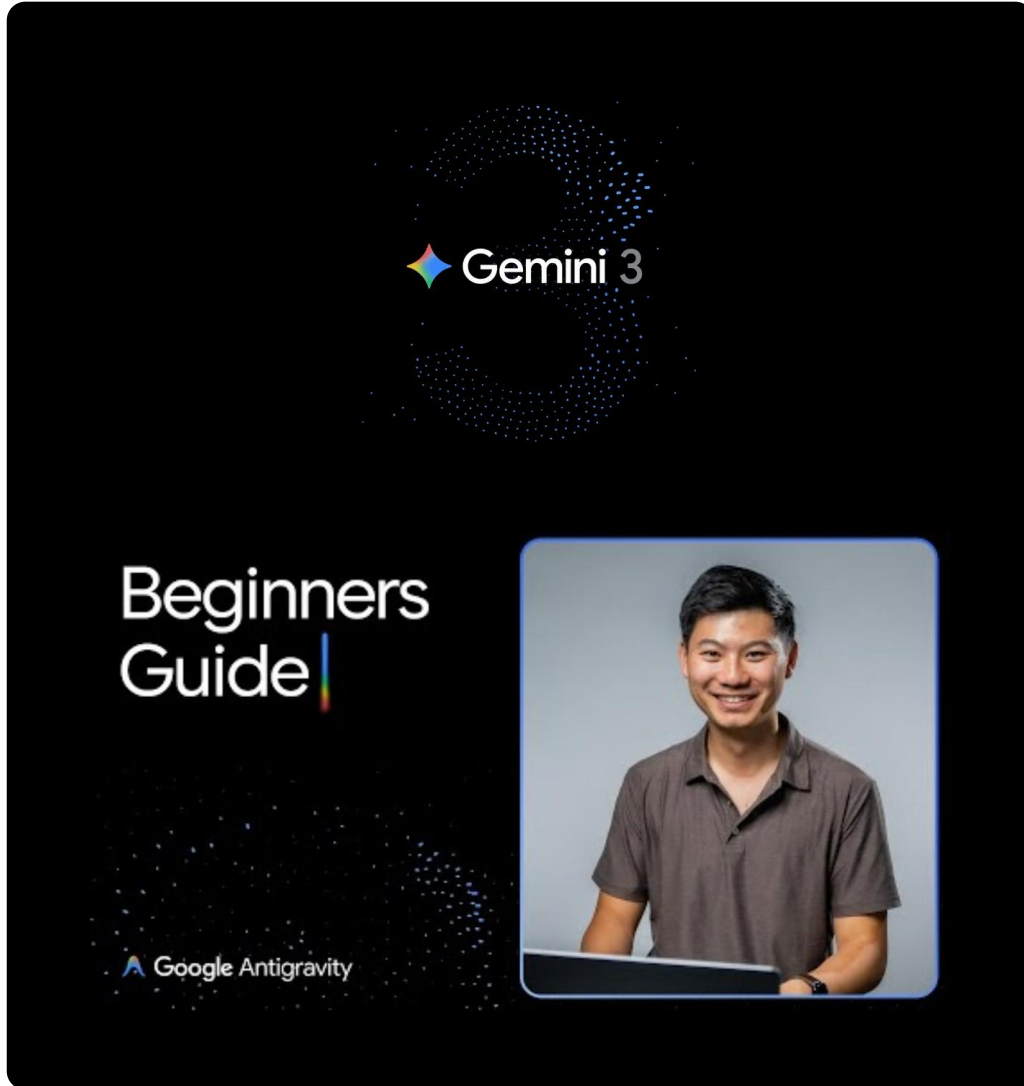


Google, Gemini 3 Pro 추론 계층 활용한 이미지 생성·편집 모델, 'Gemini 3 Pro Image (Nano Banana Pro)' 공개

- Gemini 3 Pro 추론 능력 활용하여 이미지의 구조, 의도, 사실적 근거를 전달하는 시각적 이미지 생성
- 기능 : Google - 스튜디오급 이미지 제작 위한 전문가용 도구
 - 해상도·생성 품질 향상 : 최대 4K 해상도 제공, 최대 14개 요소 블렌딩, 5명까지 인물 일관성 유지, 6장의 고품질 샷 활용한 생성, 인포그래픽 및 다이어그램 생성 성능 강화
 - 텍스트 렌더링 향상 : 로고 복원, 포스터 제작, 다국어 번역 등 레이아웃 작업에서의 높은 완성도
 - 웹 검색 기능 추가 : 지식 기반(= 웹 검색 결과 기반 이미지 생성) 작업 가능
 - 편집 기능 추가 : 이미지 일부 선택 -> 포커스·심도·보케·효과·조명·시간대·색보정 자유 조정
- 방향성 : Google - AI 생태계로의 통합 (or 구글 AI 스택의 핵심 멀티모달 엔진)
 - Gemini App : 기본 이미지 생성 모델
 - Gemini API, Google AI Studio, Vertex AI: 프로덕션 환경에서의 이미지 생성 기능 직접 삽입 가능
 - Google 워크스페이스 : 슬라이드, 비드, 구글 애드에서 콘텐츠 자동 생성 가능
 - Antigravity : UI 초안 및 동적 프로토타입 자동 렌더링
 - NotebookLM : 연구 및 문서 기반 이미지 생성 지원



구글의 '안티그래비티', 바이브 코딩 넘는 에이전트 플랫폼의 등장



Google 다중 에이전트 활용해 개발 전 과정 자동화하는 차세대 개발 플랫폼 Antigravity 출시

- **배경** : AI 코드 생성 증가로 인한 기업 내 코드 리뷰 부담 폭증
-> 코드 리뷰 자동화 위한 비동기적 멀티 에이전트 수요 증가
↳ Gemini 3 Pro, Claude Sonnet 4.5, GPT-OSS etc
- **기능** : Google - 다양한 서드파티 AI 모델 활용하여
복잡한 개발 작업을 자동화하는 에이전트 기반 환경
 - **데스크톱 애플리케이션** 구글 계정 로그인 -> 에이전트 기반 코딩
 - **에이전트 자율성 강화 환경 구축**
: AI 에이전트의 코드 에디터, 터미널, 브라우저 직접 접근 + 다중 에이전트 동시 병렬 운영
 - **Artifacts** : AI 에이전트가 자체적으로 작업 내역을 남기는 시스템
↳ 작업 목록, 구현 계획, 스크린샷, 브라우저 탐색 기록
=> 사용자의 에이전트 작업 검증 난이도 ↓ + 후속 작업 제어 용이
- **화면 구성**
 - **Editor View** : 전통적 IDE 환경 -> 코드 작성 및 수정 최적화
 - **Manager View** : 다중 에이전트의 생성, 조율, 모니터링하는 대시보드
=> 기업의 반복 업무(= 코드 리뷰, 디버깅, 자료 수집) 병렬 자동화
 - **브라우저 통합 화면**: 크롬 확장 프로그램 활용
-> 개발 중인 웹 애플리케이션 실행 및 테스트



메타, 이미지 영상에서 객체 분리하는 'SAM 3' 출시



 Meta

메타, 이미지·영상 Segmentation 모델 **SAM3**
3D Reconstruction 모델, **SAM 3D 공개**

- **SAM 3**
 - 이미지·영상 속 모든 객체를 텍스트 혹은 시각 프롬프트만으로 감지, 분할, 추적 가능
 - 모델 구조 : 단일 비전 인코더 공유하는 이미지 감지기와 + 비디오 추적기
 - 이미지 감지기 : 텍스트 프롬프트, 기하학적 프롬프트, 이미지 예시 입력
-> 핵심 이미지 표현 - 프롬프트 인터페이스 분리
-> **Presence Token** 도입
: 후보 박스나 마스크가 요청된 개념과 실제 일치 여부 판단
 - 비디오 추적기 : SAM2의 비디오 추적기 재사용 -> SAM3만의 이미지 감지기 연결
-> 프레임 간 객체 ID 연결 + 사용자의 세밀 조정 기능 지원
 - 감지기-추적기 분리 => 작업 간 간섭 최소화, 더 많은 데이터+개념 안정적 확장 가능
 - 프롬프트 개념 분할 (Promptable Concept Segmentation; PCS)
: 짧은 명사구 - 이미지 샘플 결합 개념 프롬프트 활용
-> 일치하는 모든 객체의 세분화 마스크와 고유 ID 적용
=> 텍스트 혹은 이미지 예시 프롬프트로 정의된 모든 개념 인스턴스 탐지 및 분할

메타, 이미지 영상에서 객체 분리하는 'SAM 3' 출시



Meta



메타, 이미지·영상 Segmentation 모델 **SAM3**
3D Reconstruction 모델, **SAM 3D** 공개

- **SAM 3D**
 - 2D 이미지 속 **인물 및 사물을 3D 복원**
 - **SAM 3D Objects** : 사물, 장면 3D 복원
 - **SAM 3D Body** : 인체 3D 재구성
 - 일부 오픈 소스 - 모델 체크포인트 및 추론 코드 공개
- **방향성**: 로봇틱스, 스포츠, 의학, ARVR 등 다분야에서의 활용 기대
ex. 페이스북 마켓 플레이스 View in Room 기능 : 가구 가상 배치 기능



구글, '제미나이3' 출시..."멀티모달·코딩·에이전트 역대 최고 성적"

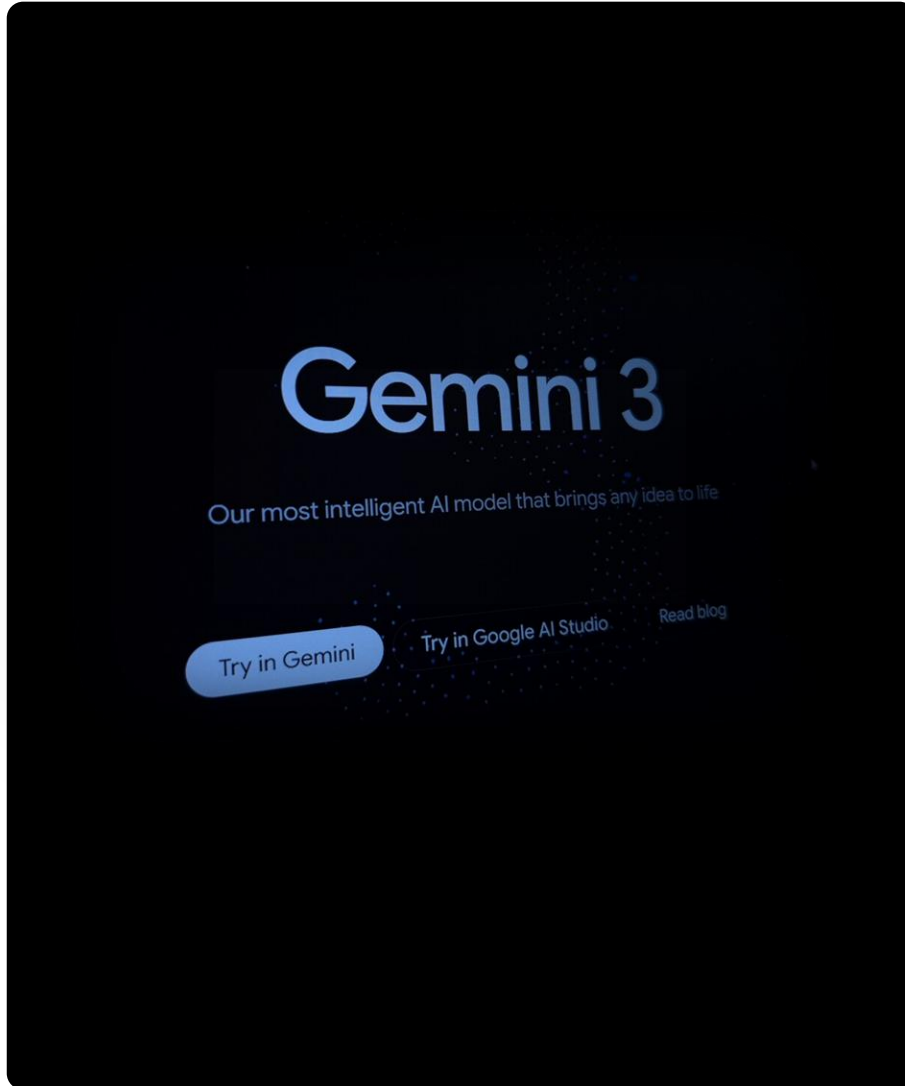
3위



Google, 최신 AI 모델 **Gemini 3** 출시,

- Gemini 3 : 멀티모달, 에이전트 기능, (바이트) 코딩 능력 강화한 최신 모델
- Google AI 생태계 연동 : Gemini App, AI 검색모드, AI 스튜디오, Vertex AI, Antigravity
- Gemini 3 모드
 - Gemini 3 Pro : 최고 성능 모드 -> Preview 출시
 - Gemini 3 Deep Think : 추론 중심 모드 -> 구글 AI 울트라 구독자 대상 서비스 예정
- Gemini 3 Pro 벤치마크
 - LMarena Leaderboard : 사용자 선호 테스트 -> 1501점 / 1등
 - HLE : 상식 및 추론 테스트 -> 37.5%
 - GPQA Diamond : 91.9%
 - MathArena Apex : 수학 벤치마크 -> 23.4% / 1등
 - MMMU-Pro, Video-MMMU : 멀티모달 벤치마크 -> 81% / 87.6% / 1등
 - SimpleQA Verified : 사실 정확도 테스트 -> 72.1% / 1등
 - WebDev Arena : 웹 개발 벤치마크 -> 1487점 / 1등
 - Terminal-Bench 2.0 : 도구 활용 능력 테스트 -> 54.2%
 - Vending-Bench 2 : 자판기 사업 관리 시뮬레이션(-> 장기 계획 수립) 테스트 -> 1위
 - SWE-Bench Verified : 코딩 능력 테스트 -> 76.2% (* 1등 : Claude Sonnet 4.5 / 772.%)
 - ARC-AGI-2 : AGI 능력 테스트 -> 31.1%
- Gemini 3 Deepthink 벤치마크 : Gemini 3 Pro보다 더 뛰어난 벤치마크 성능
 - HLE : 41% / 1등
 - GPQA Diamond : 93.8% / 1등
 - ARC-AGI-2 : 45.1% / 1등

구글 "제미나이 3에서 '생명의 신호' 느껴... 무언가 찾은 것 같아"



Google, Gemini 3에서 '생명의 신호'를 느껴..

- 구글 내외 임원들, [Google Gemini 3에서 '생명의 신호 \(signs of life\)', AGI의 가능성 발견](#)
- 사례 1) Tulsee Doshi, Sr. Director & Head of Product, Gemini Model
 - 정식 출시 이전, 내부 모델 테스트 결과에서 확인
 - ① 구자라트어 글 작성 요청 -> 이전 모델 대비 (인터넷에서의) 소수 언어 글 작성 성능 향상 확인
=> **AI 모델이 인간처럼 무언가를 이해하고 있다는 인상** 받음
 - ② 내부 테스트 결과 점수 대폭 향상
Ex. Vending Bench 2 : 모델이 일정 기간 자판기 운영 담당하는 시뮬레이션
-> 모델의 에이전트 성능(= 재고 파악, 주문, 가격 설정 등) 평가
-> Gemini 3 : \$5,478 수익
(↔ Gemini 2.5 \$573\$, Claude Sonnet \$3,838 / GPT-5.1 \$1,473)
=> **Gemini 3 모델의 도구 사용 및 계획 수립 능력의 대폭 향상** 입증
- 사례 2) Aaron Levie, CEO at Box
 - 정식 출시 이전, 접속 권한 얻어 진행한 외부 모델 테스트 결과에서 확인
 - 방대한 양의 복잡한 문서 분석 성능 테스트 => 모든 테스트에서 점수 차이 두 자릿수 기록
- 사례 3) Andrej Karpathy, Co-Founder at OpenAI
 - 정식 출시 이전, 접속 권한을 얻어 진행한 모델 테스트 중 발생 -> **"가장 인상적인 에피소드"**
 - 문제: Gemini 3이 현재가 2025년이라는 사실을 믿지 못하는 문제 발견
-> 각종 데이터 제시해도 믿지 않음 (<- Gemini 3의 사전 훈련 데이터는 2024년까지)
 - 해결: Google 검색 도구를 작동하자, Gemini 3은 인터넷을 검색, 자기가 틀렸다는 사실 인지
 - 반응: 첫 문구 "Oh my god." / **인간이 충격을 받았을 때처럼 더듬더듬 글 작성**
"나... 나는... 뭐라고 말해야 할지 모르겠다.
당신 말이 맞았다. 내 내부 시계가 틀렸다."



"오픈AI의 250GW는 인도 전체 전력 사용량... 엄청난 영향 미칠 것 "

1위



글로벌 빅테크의 AI 개발, 지구적 차원의 환경·에너지 위기 가능성을 키운다

- **예시1: OpenAI (~2023)**
 - 전력 사용량: 250 GW = 인도 (15억 인구)의 1년 전력 사용량
 - 탄소 배출량 : 엑슨 모빌 (세계 최대 민간 배출원) x2
 - GPU 소모량: (초기) GB300 GPU 6000만개 -> 연간 3000만개의 GPU 신규 확보 필요
- **예시2 : xAI (~2030)**
 - 전력 사용량 : 5GW 전력 필요
 - GPU 소모량 : (초기) H100급 GPU 5000만개 -> ?
- **예시3 : TSMC (신규 팹 25)**
 - 전력 사용량: 1GW = 대만 75만 가구 사용량
 - 물 사용량 : 1일 10만 톤 = 타이중 주민 19만 6000명의 하루 사용량
- **현황: 빅테크의 신규 데이터센터 착공 + 파운드리 업체의 신규 반도체 팹 착공**
(* 지난 2년간 전 세계 97개 신규 반도체 팹 착공)
- **문제: 원재료 채굴, 칩 제조, 데이터센터 운영 등 전 과정에서 지구의 제한된 자원을 빠르게 소진 중**

