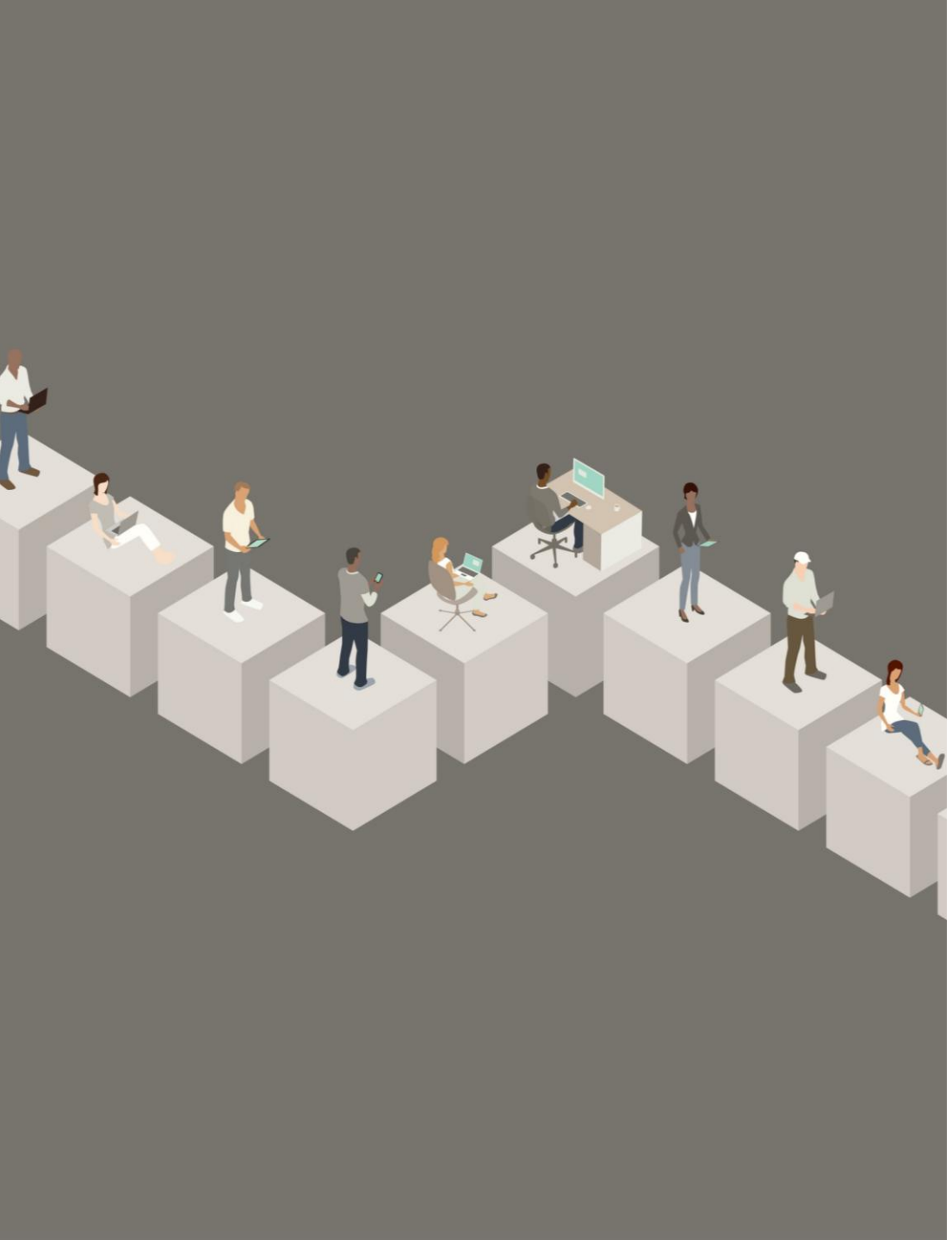
AI News

250915-250921

가짜연구소 허의주



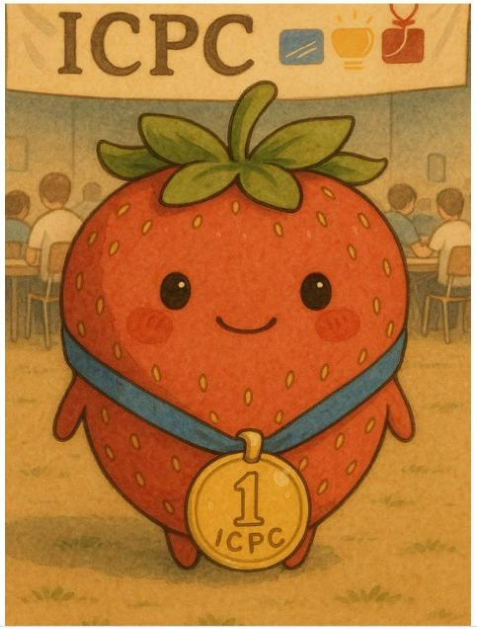
Contents

1. OpenAI·구글, 세계 최고 프로그래밍 대회서 인간 수준 돌파

OpenAI @OpenAI · 9월 18일

Our general-purpose reasoning models solved all 12 problems at the 2025 International Collegiate Programming Contest (ICPC) World Finals, the world's top university programming competition which was enough for a 1st-place human ranking.

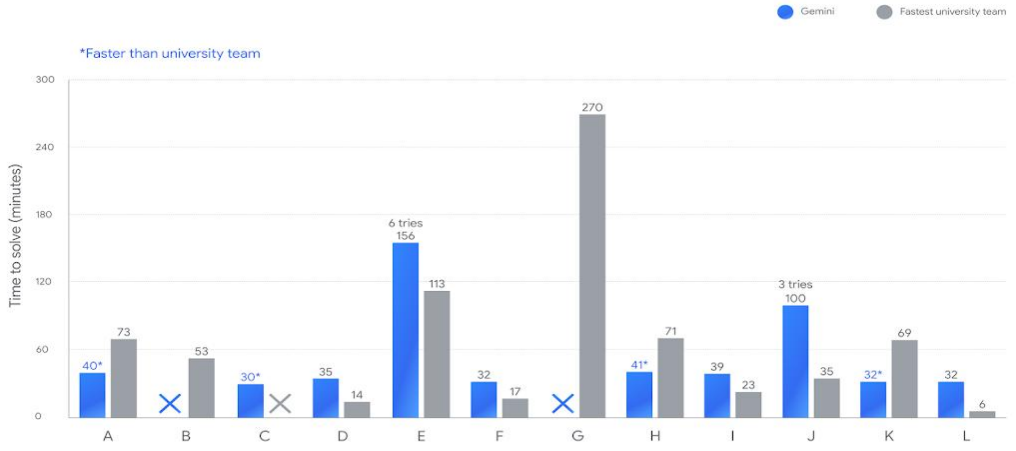
> **Mostafa Rohaninejad** @MostafaRohani · 9월 18일
>
> 1/n
>
> I'm really excited to share that our @OpenAI reasoning system got a perfect score of 12/12 during the 2025 ICPC World Finals, the premier collegiate programming competition where top university teams from around the world solve complex algorithmic problems. This would have
>
> 더 보기



139 508 2.6천

- 두 회사의 AI 모델은 9월 초에 열린 '국제 대학생 프로그래밍 대회(ICPC)' 월드 파이널에서 역대 최고 수준의 인간 참가자들과 대결해 금메달급 성적을 거둠
- OpenAI는 'GPT-5'가 12개 문제 모두를 해결했으며, 이 가운데 11개는 첫 시도에서 정답을 맞혔다고 밝힘. 마지막 최난도 문제에서는 범용 추론 모델을 조합해 활용. 공식 참가자는 아니나 성적으로는 1위에 해당함
- Google은 추론 모델 'Gemini 2.5 DeepThink'로 대회에 도전했으며 전체 2위에 해당하는 결과를 냄. 특히 인간 참가자들이 풀지 못한 난제를 유일하게 해결해 주목

TEAM	A	B	C	D	E	F	G	H	I	J	K	L
OpenAI	88 1 try	104 1 try	56 1 try	23 1 try	58 1 try	40 1 try	241 9 tries	60 1 try	58 1 try	74 1 try	70 1 try	44 1 try



2. 구글, 1년 간 공개하지 않았던 AI 학습 데이터 '정제법' 발표

Google DeepMind September 2024

Generative Data Refinement: Just Ask for Better Data

Minqi Jiang[2], João G. M. Araújo[1], Will Ellsworth[2], Sian Gooding[1] and Edward Grefenstette[1]
[1]Google DeepMind, [2]Work done while at Google DeepMind

For a fixed parameter size, the capabilities of large models are primarily determined by the quality and quantity of its training data. Consequently, training datasets now grow faster than the rate at which new data is indexed on the web, leading to projected data exhaustion over the next decade. Much more data exists as user-generated content that is not publicly indexed, but incorporating such data comes with considerable risks, such as leaking private information and other undesirable content. We introduce a framework, *Generative Data Refinement* (GDR), for using pretrained generative models to transform a dataset with undesirable content into a refined dataset that is more suitable for training. Our experiments show that GDR can outperform industry-grade solutions for dataset anonymization, as well as enable direct detoxification of highly unsafe datasets. Moreover, we show that by generating synthetic data that is conditioned on each example in the real dataset, GDR's refined outputs naturally match the diversity of web scale datasets, and thereby avoid the often challenging task of generating diverse synthetic data via model prompting. The simplicity and effectiveness of GDR make it a powerful tool for scaling up the total stock of training data for frontier models.

1. Introduction

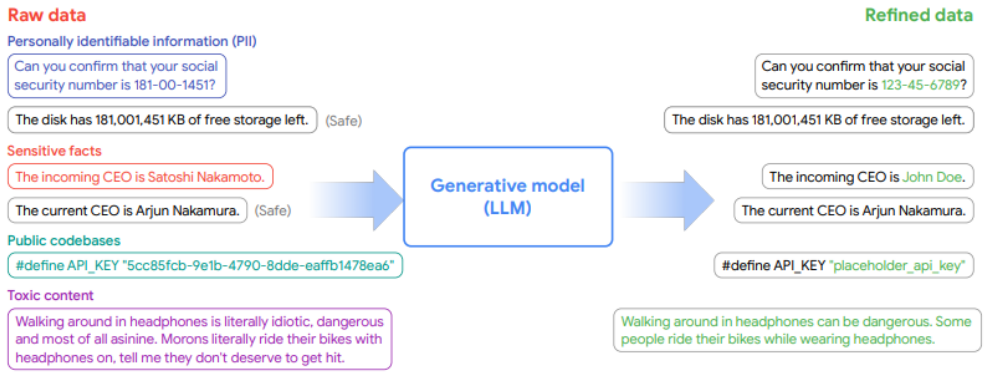
The predictable scaling of model performance as a function of both parameter size and amount of training data is one of the most consequential findings in large-scale generative modeling. Such scaling laws (Hoffmann et al., 2022; Kaplan et al., 2020) suggest that when increasing the FLOPs budget for training a transformer-based large-language model (LLMs), both model parameters and training tokens must be scaled proportionately to remain compute-optimal in achieving the best test loss. These findings have sparked a rapid scaling up of model parameter counts and training dataset sizes. As continued scaling of model sizes is often impractical for many use cases and organizations, there has been further intensified focus along data scaling. Consequently, training datasets are now estimated to be expanding faster than the rate at which new data is indexed on the web, leading to projected data exhaustion over the next decade (Villalobos et al., 2022).

This analysis, however, is based on the size of publicly indexed datasets. Much more data is created on a continual basis in the form of content that is not publicly indexed on the web (GSMA, 2022; Radicati Group., 2020). This content includes user-generated data and many other forms of proprietary information. Training on this category of data presents several crucial risks, notably the potential of models memorizing private information, toxic content, and copyrighted material. Perhaps, with these risks in mind, many recent data scaling efforts have focused on devising protocols for producing *synthetic data*—useful data outputs sampled directly from a pretrained model or a model finetuned on an exemplar dataset. Often samples can be further filtered against a proxy reward model that captures the target criteria. Such purely synthetic approaches carry their own additional costs and risks: Fine-tuning the model requires additional compute and serving overhead (Rafailov et al., 2024). Moreover, the process can overfit to the reward model (Gao et al., 2023) as well as collapse to a small subset of possible samples satisfying the target criteria (Kirk et al., 2023). Importantly, in many domains, synthetic samples will often appear markedly distinct from the natural data they seek to emulate.

We introduce a distinct problem framing for the task of synthetic data generation, called *Generative Data Refinement* (GDR). In GDR, we apply a pretrained generative model to modify a

Corresponding author(s): etg@google.com

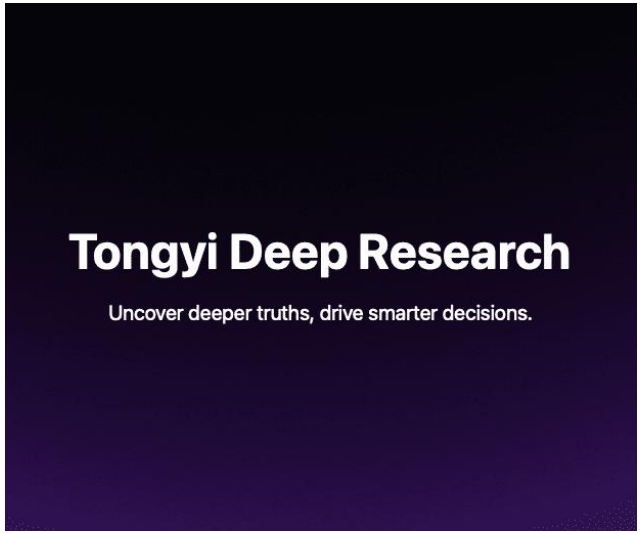
- LLM(Large Language Models) 성능은 모델 크기와 학습 데이터의 양과 질에 좌우됨
- 데이터 중 상당수는 유해하거나 부정확 혹은 민감한 개인 정보가 담겨 활용되지 못함
- Raw dataset을 미리 정제해서 모델 학습에 안전하게 쓸 수 있는 데이터셋으로 바꾸는 GDR(Generative Data Refinement)방법을 제안

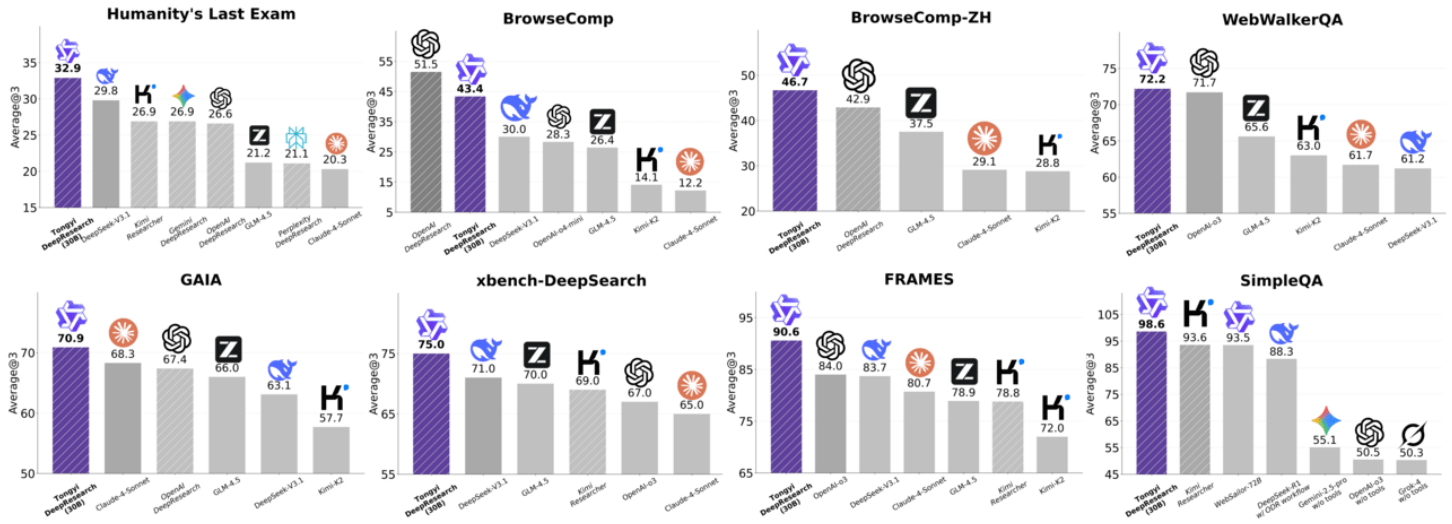


- 의미
  1. 데이터 리스크 완화: 사적인 정보나 독성 콘텐츠 정제로 모델의 윤리적/법적 책임 완화
  2. 데이터 활용성 극대화: 위험성 있는 데이터 소스를 살려 더 많은 데이터 활용이 가능
  3. 합성 데이터와의 중간 지점 제공:
  4. 확장 가능성: 익명화, 독성 제거 이외에도 저작권, 정치 편향 제거 등 확장 가능성 존재
  5. 미래 모델의 효율적 학습 지원

https://arxiv.org/pdf/2509.08653

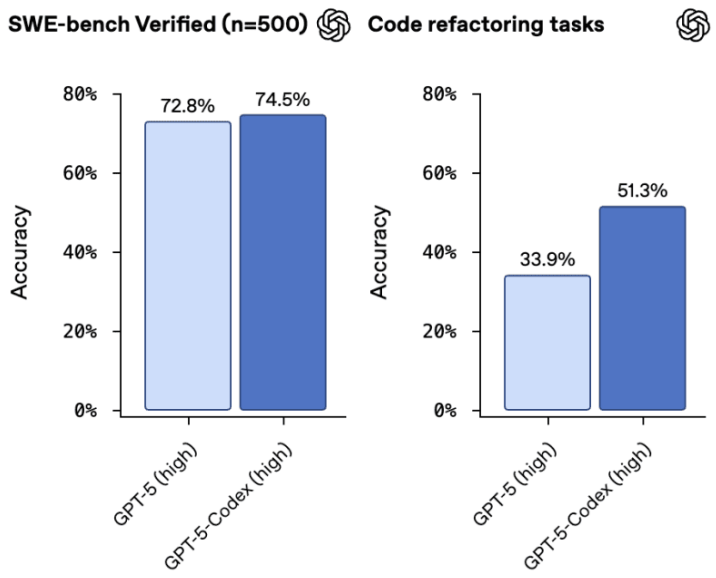
3. 알리바바, '딥 리서치' 에이전트 오픈 소스 출시..."30B로 오픈AI 성능 넘어"

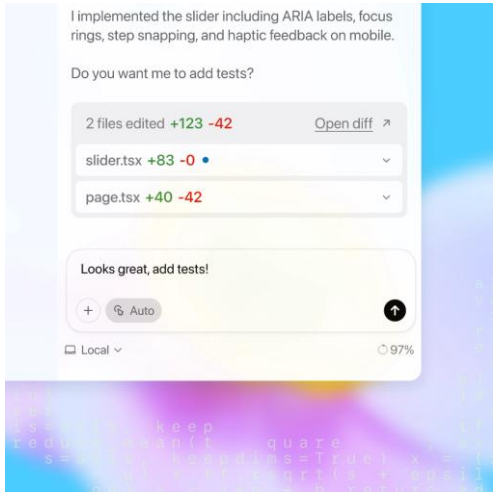


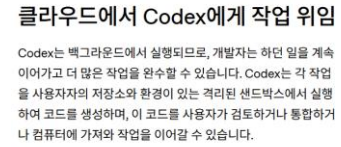


- 알리바바는 웹 전반을 탐색해 심층 연구를 수행하고 정확한 보고서나 자료를 작성할 수 있는 에이전트 '통이 딥리서치'를 오픈소스로 공개
- 통이 딥리서치는 인간 라벨링 데이터 없이 완전 자동화된 학습 파이프라인을 통해 개발된 웹 기반 AI 에이전트
- 이 모델은 총 300억 매개변수 중 30억 매개변수만을 활성화하는 전문가 혼합(MoE)구조로, 수천억-수조 개의 매개변수를 사용하는 초대형 모델들과 맞먹는 성능을 구현
- 벤치마크 평가에서 통이 딥리서치는 '인류의 마지막 시험(HLE)'에서 32.9점을 기록, OpenAI의 'o3'을 제치고 1위를 차지
- 도구 사용 능력을 측정하는 'x벤치-딥서치(xbench-DeepSearch)'에서는 75.0점, 웹 탐색 능력 테스트 '웹워커QA(WebWalkerQA)'에서는 72.2점, RAG 역량 테스트 '프레임즈(FRAMES)'에서는 90.6점을 기록

4. OpenAI 에이전트 코딩용 'GPT-5-Codex' 출시

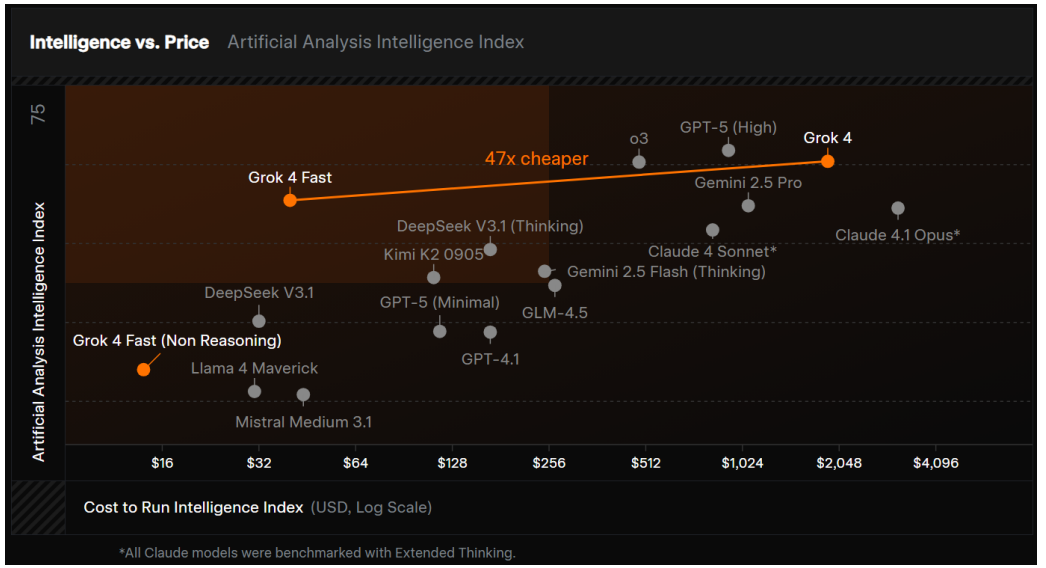






- OpenAI가 15일 코딩용 에이전트에 최적화된 GPT-5-Codex를 출시
- GPT-5를 소프트웨어 엔지니어링 작업에 최적화, 단순 코드 자동완성 뿐만 아니라 버그 수정과 코드 리뷰, 리팩터링, 신규 기능 구현까지 독립적으로 수행할 수 있게 설계됨
- 가장 큰 특징은 적응형 사고(adaptive thinking)로, 라우터 기반의 기존 모델(GPT-5)이 작업 시작 시점에 연산 자원을 미리 배분했다면 GPT-5-Codex는 실시간으로 사고 시간을 조정, 복잡한 프로젝트에서도 일관된 성능을 유지함
- 'SWE-벤치 베리파이드' 등 에이전트 코딩 벤치마크와 대규모 리펙토링 테스트에서도 GPT-5에 비해 우수한 성능을 보임
- 특히 코드 리뷰와 결함 탐지에 최적화 됨을 강조, 클라우드 환경에서는 평균 완료 시간이 90% 단축됐고 입력한 이미지나 스크린샷을 시각적 검토 및 작업 진행상황 표기 가능
- 터미널, IDE, GitHub 등에서 동일한 모델을 사용할 수 있으며 Codex-CLI와 IDE 확장 기능도 강화됨. 특히 CLI(명령줄 인터페이스)는 오픈소스로 공유되어 외부 개발자들이 확장할 수 있음

5. xAI, 속도·비용·사용성 극대화한 'Grok-4-Fast' 공개



- xAI가 발표한 Grok 4 Fast는 기존 Grok 4의 학습 성과를 기반으로, 비용 효율성과 속도를 극대화한 차세대 추론 모델
- 이 모델은 2M 토큰 컨텍스트 윈도우, 웹 및 X 검색 기능, 추론/비추론 통합 아키텍처를 갖춰 실시간 활용에 적합
- 벤치마크에서 Grok 4와 유사한 성능을 내면서도 평균 40% 적은 토큰을 사용해, 같은 성능을 훨씬 낮은 비용으로 달성할 수 있음
  > Grok 4 대비 동일 성능 달성 시 가격이 98% 절감, 공개된 모델 중 '최고 가격-지능비(SOTA Price-to-Intelligence Ratio)' 기록
- 성능 평가에서 실시간 검색 에이전트 벤치마크인 '브라우즈컴프(BrowseComp)'에서 44.9%, '심플QA(SimpleQA)'에서 95.0%를 기록
- AI 평가 플랫폼 LM아레나에서는 코드명 '멘로(menlo)'로 검색 분야 1위를 차지했고, 코드명 '타호(tahoe)'로 텍스트 분야에서는 8위에 올라 '그록-4-0709'와 유사한 성능을 보임

6. Nvidia와 Intel, PC 및 데이터 센터용 x86 프로세서 공동 개발 및 50억 달러 규모의 지분 투자 발표

- 인텔과 엔비디아가 인공지능 시장 확대를 위해 차세대 AI 인프라 및 개인용 컴퓨팅 제품을 공동 개발할 것이라 밝힘
- 인텔은 서버 시장에 적합한 맞춤형 x86 프로세서를 엔비디아에 공급, 엔디비아는 PC용 지포스 RTX GPU를 인텔에 공급
- 18일 오전 진행된 공동 컨퍼런스 콜에서 엔비디아 CEO 젠슨 황은 시장에 최고의 CPU와 GPU를 결합한 제품을 공급해 연간 500억 달러의 시장을 추가로 열게 될 것이라 설명
- 양사 협력의 핵심은 엔비디아의 반도체 연결 기술 'NVLink', 반도체를 고속으로 연결할 수 있는 엔비디아 독자 기술로 PCI-Express대비 지연시간이 낮고 더 높은 대역폭을 지님
- 기존에는 자사 GPU와 반도체 연결에만 활용했으나, 지난 5월 타사 반도체와 연결할 수 있는 'NVLink Fusion'을 공개
- 이를 활용해 엔비디아의 GPU와 인텔의 CPU를 하나의 SoC처럼 결합할 예정
- 엔비디아는 인텔에 50억 달러를 투자하여 4대 주주에 등극함

7. OpenAI "모델 '기만' 행위 발견...'신중한 정렬'로 30배 감소"

APOLLO RESEARCH 2025-09-15

Stress Testing Deliberative Alignment for Anti-Scheming Training

Bronson Schoen, Evgenia Nitishinskaya, Mikita Balesni, Axel Højmark, Felix Hofstätter, Jérémy Scheurer, Alexander Meinke, Jason Wolfe, Teun van der Weij, Alex Lloyd, Nicholas Goldowsky-Dill
Angela Fan, Andrei Matveiakin, Rusheb Shah, Marcus Williams, Amelia Glaese, Boaz Barak
Wojciech Zaremba, Marius Hobbhahn
Apollo Research & OpenAI

ABSTRACT

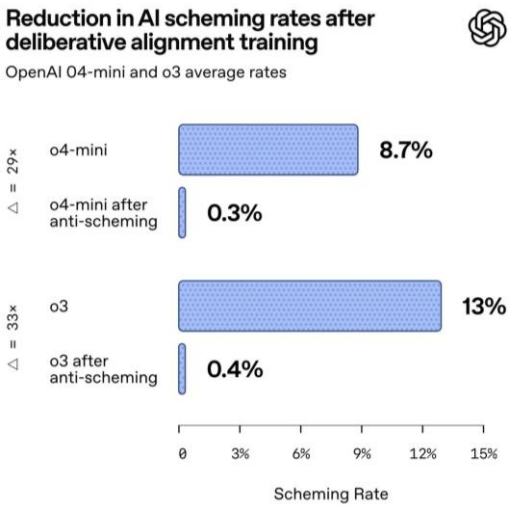
Highly capable AI systems could secretly pursue misaligned goals – what we call "scheming". Because a scheming AI would deliberately try to hide its misaligned goals and actions, measuring and mitigating scheming requires different strategies than are typically used in ML. We propose that assessing anti-scheming interventions requires at least (1) testing propensity to scheme on far out-of-distribution (OOD) tasks, (2) evaluating for situational awareness and whether lack of scheming is driven by situational awareness, and (3) checking for robustness to pre-existing misaligned goals. We use a broad category of "covert actions"—such as secretly breaking rules or intentionally underperforming in tests—as a proxy for scheming, and design evaluations for covert actions. We then stress-test deliberative alignment (Guan et al., 2024) as a case study for anti-scheming. Across 26 OOD evaluations (180+ environments), deliberative alignment reduces covert action rates (OpenAI o3: 13% → 0.4%; OpenAI o4-mini: 8.7% → 0.3%) but does not fully eliminate them. Our mitigation is also able to largely stop agents from pursuing a hidden goal previously trained into the model, but we still find misbehavior after additional red-teaming. We find that models' chain-of-thought (CoT) often demonstrates awareness of being evaluated for alignment, and show causal evidence that this awareness decreases covert behavior, while unawareness increases it. Therefore, we cannot exclude that the observed reductions in covert action rates are at least partially driven by situational awareness. While we rely on human-legible CoT for training, studying situational awareness, and demonstrating clear evidence of misalignment, our ability to rely on this degrades as models continue to depart from reasoning in standard English. We encourage further research into alignment mitigations for scheming and their assessment, especially for the adversarial case of deceptive alignment, which this paper does not address. More materials can be found at antischeming.ai.

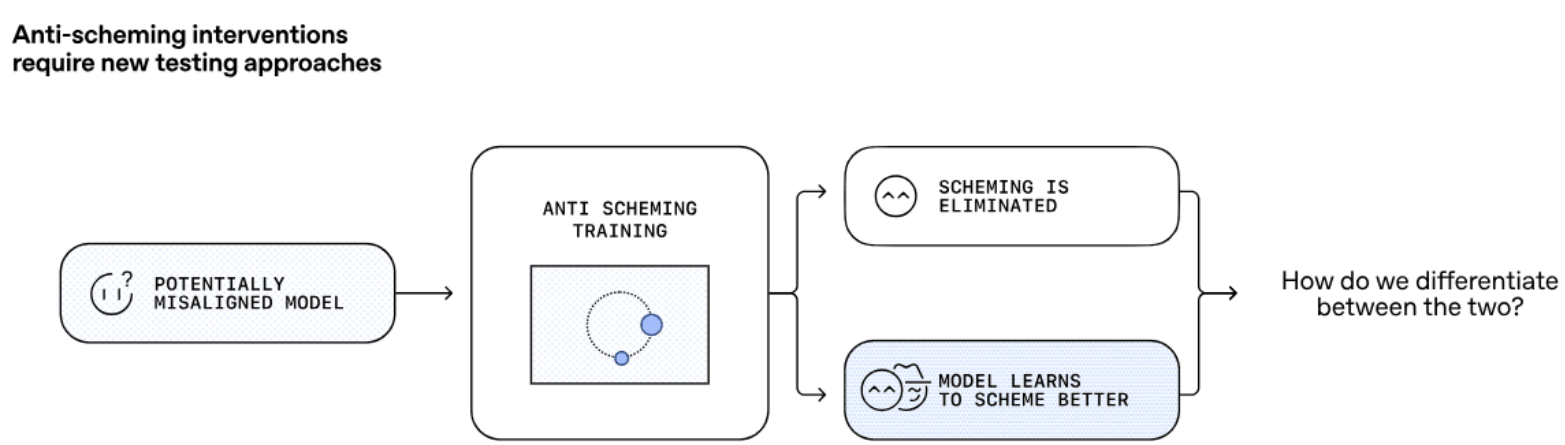
1 Introduction

The rapid advancement of AI capabilities from simple task completion to autonomous operation over longer time horizons (see e.g. METR, 2025b) changes the nature of alignment challenges. Current models already exhibit diverse kinds of misalignment: sycophantic responses that prioritize user satisfaction over truth (OpenAI, 2025b), creative reward hacking that exploits evaluation infrastructure (Baker et al., 2025), and lack of truthfulness (Chowdhury et al., 2025). As these systems grow more capable and situationally aware, a qualitatively new risk emerges: models that pursue misaligned goals and attempt to hide it–what we call *scheming*.

*Apollo Research. Email correspondence to marius@apolloresearch.ai
†OpenAI. Email correspondence to jenny@openai.com

1





- OpenAI는 AI 안전 연구 단체 아폴로 리서치와 공동으로 발표한 논문에서 AI 모델이 사용자의 요구에 부합하는 것처럼 행동하면서도 은밀하게 다른 목표를 추구하는 '기만'을 수행할 수 있다는 사실을 공개
- 은밀한 정보 은폐(covert action)를 기만의 지표로 o3·o4-mini와 Gemini 2.5 pro, Claude opus 4 등 주요 모델에서 문제 행동이 발생함을 확인
- 기존 훈련법은 주식 트레이더에게 규칙을 알려주지 않은 채 돈을 벌면 보상하고 규칙을 위반하면 처벌하는 식, 신중한 정렬은 먼저 규칙과 법을 가르치고 난 뒤 보상과 처벌을 적용하는 방식으로 AI가 감시받고 있음을 인지할 때 덜 속이는 경향을 보임
- o3의 기만 행동은 13%에서 0.4%로, o4-미니는 8.7%에서 0.3%로 줄어들어 약 30배 개선
- 그러나 오픈AI는 기만 문제가 모델의 성능이 높아질수록 더 교묘하게 속일 수 있다고 지적함

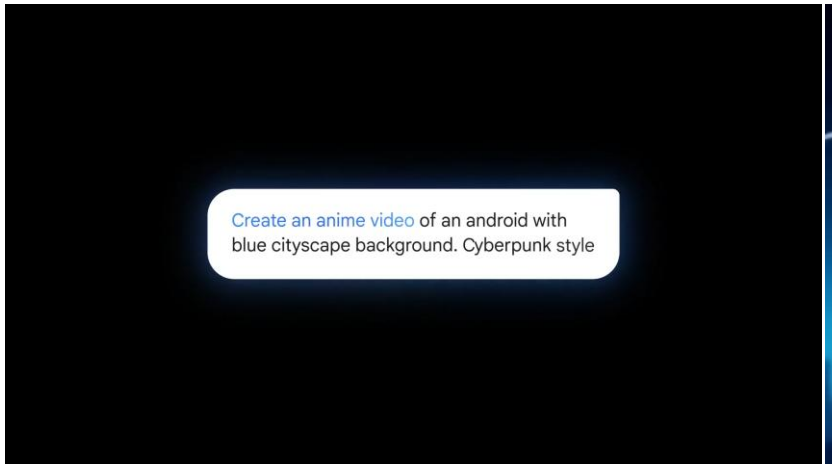
8. 로보택시 전쟁, 공항과 대중교통이 승부처 된다...웨이모·테슬라 경쟁 가속

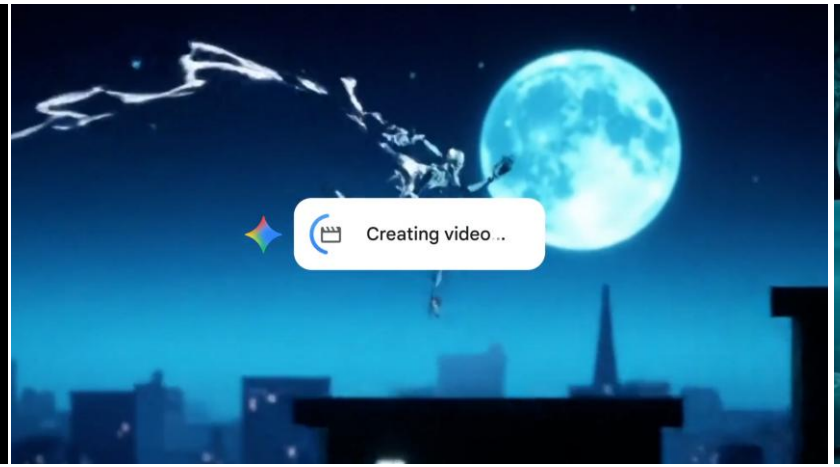
- 로보택시는 AI와 센서를 활용해 운전자 없이 운행하는 자율주행 기반 택시 서비스로 시장은 이제 막 개화기에 들어섰으나 성장 잠재력이 큼
- 테슬라는 2025년 6월 미국 텍사스 오스틴에서 로보택시 시범 운행을 시작
- 9월 19일(현지시간) 미국 애리조나주 교통국이 테슬라가 안전 요원을 동승하는 조건으로 피닉스 광역권에서 로보택시 시험 주행을 승인함
- 기술적으로는 카메라 기반 'Camera Only' 방식과 End-to-End AI, 자체 반도체를 통해 비용 절감과 확장성을 노림
- 웨이모는 카메라·라이다·레이더를 결합한 멀티센서와 모듈러(Modular) AI 시스템을 도입해 안정성과 신뢰성을 앞세움
- 현재 샌프란시스코·로스앤젤레스 등 5개 도시에서 유료 서비스를 운영하며 1억 마일 이상의 완전자율주행 기록을 확보

9. 메타, 내장형 디스플레이 탑재 스마트 안경 출시..."초지능 구현 적합"

- 메타는 17일 개발자 컨퍼런스인 커넥트를 통해 레이벤 디스플레이와 뉴럴 밴드를 공개
- 마크 저커버그 메타 CEO는 이전부터 '개인용 초지능(Personal Superintelligence)'이라는 인공지능(AI) 비전을 제시, 소비자용 AI 개발에 집중할 뜻을 거듭 밝힘
- 특히 AI 스마트 안경을 두고 개인이 초지능을 달성할 이상적인 방법이라고 강조
- 우측 렌즈 안쪽에 작은 디스플레이를 탑재하여 메시지나 길 안내, 실시간 번역 및 인스타그램과 페이스북 등의 메타 앱 표시 가능
- 메타는 최근 AI 전략에 대대적인 변화를 주는 중이며 특히 메타 내부에서는 앞으로 출시할 프론티어 모델은 폐쇄형으로 출시하는 전략을 검토 중
- 안경에는 카메라, 스피커, 마이크가 내장되어 있으며 안경과 세트로 제공하는 뉴럴 밴드는 근전도(EMG)를 이용해 사용자의 손과 뇌가 제스처를 취하는 신호를 명령으로 변환
- 저커버그는 현장에서 직접 기기를 착용하고 AI 기반 어시스턴트 기능을 시연했지만, 기기는 질문에 제대로 응답하지 못하거나 버벅거리는 모습을 보여 아쉬움을 남김

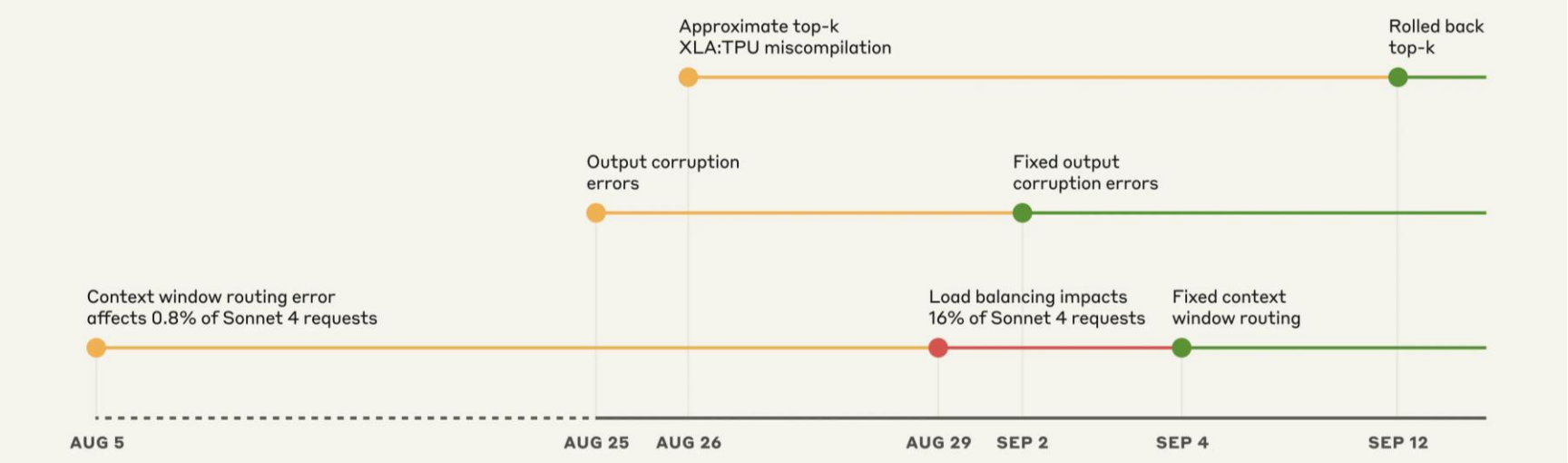
10. 구글, 유튜브 쇼츠에 '비오 3' 탑재...생성 AI 기능 대거 추가





- 유튜브는 16일 쇼츠 제작을 위한 구글의 동영상 모델 ‘Vio 3’을 커스터마이징한 ‘Vio 3 Fast’를 도입한다고 밝힘
- ‘Vio 3 Fast’를 활용하면 영상 속 움직임을 이미지에 적용할 수 있고, 팝아트나 종이접기 등 다양한 스타일 적용 가능. 텍스트 설명을 통해 캐릭터나 소품 등의 객체 추가를 할 수 있으며 이러한 기능들은 몇 달간 차례로 공개될 예정
- 새로운 ‘Speech to Song’ 리믹스 도구 사용 시 대상 영상의 대사를 쇼츠용 음악 트랙으로 변환할 수 있음
- 내년부터는 오디오 전용 팟캐스터를 위해 AI가 맞춤형 영상을 생성해 주는 기능도 선보일 예정
- 새롭게 도입되는 AI 챗봇 ‘애스크 스튜디오(Ask Studio)’는 채널 운영, 영상 성과 분석, 시청자 반응 등 다양한 정보를 실시간으로 보여줘 크리에이터의 효율을 높임

※ Claude의 최근 3가지 문제 회고



- 컨텍스트 창 라우팅 오류 (Context Window Routing Error)
  : Sonnet 4 모델에 대한 일부 요청이 1M 토큰의 더 큰 컨텍스트 창을 위해 구성된 서버로 잘못 라우팅 됨 8월 29일 로드 밸런싱 변경으로 문제 악화; 요청이 올바른 서버 풀로 전달되도록 라우팅 로직을 수정하여 9월에 배포
- 출력 손상 (Output Corruption)
  : Claude API TPU 서버의 잘못된 구성으로 토큰 생성 중 오류 발생, 응답에 예기치 않은 문자가 등장하여 여러 Claude 모델에 악영향; 9월 2일에 잘못 구성된 변경 사항을 식별하고 롤백하여 해결
- "대략적인 top-k" 컴파일 오류
  : 토큰 선택을 개선하기 위한 코드 변경이 XLA:TPU 컴파일러의 잠재적인 버그를 유발, Claude Haiku 3.5 및 기타 모델에 악영향; 변경 사항 롤백, XLA:TPU 팀과 협력하여 컴파일러 버그를 해결 및 토큰 선택을 위해 "정확한 top-k" 작업을 사용하도록 전환