

AI-ის სამყარო დღეს

პარასკევის ნომერი: 2026 წლის 8-9 ოქტომბრის ამბები

პოლიტიკა და რეგულირება

Anthropic წესებს გადაწერს. 2026 წლის გამოყენების პოლიტიკა, რომელიც 12 ნოემბრიდან შედის ძალაში, ყალბ ანგარიშებსა და გაყალბებულ ახალ ამბებზე აკრძალვებს მატყუარა კამპანიების ერთ განყოფილებაში აერთიანებს და პერსონალიზებული საარჩევნო თარგეთინგის საყოველთაო აკრძალვას ხსნის. დაზიანების მიყენების უნარის მქონე აპარატურასთან დაკავშირებულ მოდელებს ახლა გაჩერების ღილაკი და უსაფრთხო მდგომარეობა სჭირდებათ. (იხ. სარედაქციო.) [anthropic.com]

OpenAI-მ პირველი მე-5 კატეგორიის გავლენის ოპერაცია გამოაშკარავა. რუსული წარმოშობის ქსელმა „Dark Clark“ ლათინოამერიკელი თანამშრომლები ყალბ ანალიტიკურ ცენტრში ჩაითრია; ირანულმა „Bogus Bylines“ შვიდი გამოგონილი ჟურნალისტის სახელით თითქმის 100 სტატია გამოაქვეყნა. ორივე ChatGPT-დან დაიბლოკა. [openai.com]

გათავისუფლებული მკვლევრები პასუხობენ. Jasmine Wang, Tomek Korbak და Mikita Balesni OpenAI-ის ბრალდებებს უარყოფენ და აფრთხილებენ, რომ გარე თანამშრომლობა შიშის ქვეშ მოექცა. OpenAI ამბობს, რომ გამოძიებამ „დარღვევების ნიმუში“ გამოავლინა. (იხ. სარედაქციო.) [techcrunch.com]

USA Today უჩივის OpenAI-ს 250 მლნ დოლარზე მეტს და ამტკიცებს, რომ „ასობით ათასი“ სტატია დაკოპირდა. სარჩელი უერთდება საქმეებს, რომლებშიც უკვე არიან *New York Times*, Ziff Davis და თითქმის 400 ადგილობრივი გაზეთი. [theverge.com]

ჩინეთი ტემპს არ შეანელებს. SemiAnalysis-მა ცხრა წამყვანი ჩინელი დეველოპერის 857 გამოშვება შეისწავლა: მხოლოდ 3.6%-ს

გამოუქვეყნებია ოდესმე უსაფრთხოების შედეგი, გამოშვებისას კი 1.1%-ს. პეკინის წესები აპლიკაციის დონეზე მრავლდება და არა ფრონტირზე. [newsletter.semanalysis.com]

ბიზნესი

OpenAI-ის შემოსავალი, თავიდან დათვლილი. *Financial Times*-ის ცნობით, OpenAI-მ ინვესტორებს უთხრა, რომ წლიური შემოსავალი „50 მილიარდ დოლარს უახლოვდება“ — დაახლოებით 20 მილიარდით ნაკლები გავრცელებულ 70-მილიარდიან ციფრზე, რომელიც Anthropic-ის დათვლის მეთოდს ბაძავდა. IPO-ზე საუბარი 2027 წლის დასაწყისამდე გადაიწია. [techcrunch.com]

Arena, მოდელების რეიტინგი, 3.1 მილიარდი ღირს Lightspeed-ისა და Khosla-ს ხელმძღვანელობით მოზიდული 200-მილიონიანი B სერიის შემდეგ — თითქმის ორჯერ მეტი, ვიდრე იანვარში. მის ახალ alignment-რეიტინგში, რომელიც მატყუარა დასრულებებს აფასებს, ამჟამად OpenAI-ის მოდელები ლიდერობენ. [techcrunch.com]

Manus-მა 500 მლნ დოლარზე მეტი მოიზიდა Boyu-სა და IDG-ის ხელმძღვანელობით — პირველი რაუნდი მას შემდეგ, რაც პეკინმა Meta-ს ~2-მილიარდიანი გარიგების გაუქმება აიძულა. შეფასება არ გამჟღავნებულა. [techcrunch.com]

Oracle OpenAI-ზე დიდად დებს: 130 ათასი ChatGPT-ისა და 95 ათასზე მეტი Codex-ის ლიცენზია. OpenAI ამტკიცებს, რომ კადრების მოზიდვის კვლევის დრო 98%-ით, დღეებიდან 15-20 წუთამდე შემცირდა. [openai.com]

LegalOn-მა Codex-ის ხარჯი განახევრა. სამუშაოს სირთულის მიხედვით GPT-6 Luna-ს, Sol-სა და Astra-ს შორის განაწილებამ და ბიუჯეტის ლიმიტებმა სავარაუდო დღიური ხარჯი ~65%-ით შეამცირა. [openai.com]

ლაბორატორიები

Anthropic-ის Cyber Mission ფრონტირულ Claude-ს, ადგილზე მომუშავე ინჟინრებსა და საფრთხეების კვლევას CrowdStrike-ის, Palo Alto Networks-ის, Rockwell-ისა და სხვების გვერდით აყენებს ელექტროქსელების, წყალმომარაგებისა და ტრანსპორტის დასაცავად. უფასო OSS Scanner ღია კოდის პროექტებს მოდელის მიერ შექმნილ, ადამიანის მიერ შეუმოწმებელ ანგარიშებს უგზავნის. (იხ. სარედაქციო.) [anthropic.com]

150 მლნ დოლარი ფედერალური

მეცნიერებისთვის. Anthropic Claude-სა და API-კრედიტებს თეთრი სახლის Genesis Mission-ის ფარგლებში აშშ-ის 15-ზე მეტი უწყების, მათ შორის NASA-ს, NIH-ისა და NSF-ის, რამდენიმე ასეულ პროექტს მიაწვდის. [anthropic.com]

ცის სრული ულტრაიისფერი რუკა.

ასტროფიზიკოსმა Brice Ménard-მა Claude Science-ის აგენტებით GALEX-ის დატოვებული ხარვეზები შეავსო; საკონტროლო მონაცემებზე სიზუსტე ~10%-ის ფარგლებშია. დაკვირვების დარჩენილი არტეფაქტები ადამიანმა შენიშნა და არა აგენტებმა. [anthropic.com]

პროდუქტები

Gemini კოლეგა ხდება. Google-ის კორპორატიული აგენტი იღებს საკუთარ Workspace-იდენტობას, ელფოსტასა და აუდიტის ჟურნალს და უკავშირდება Slack-ს, Jira-ს, M365-ს, მონაცემთა საცავებსა და MCP სერვერებს. პირველად — ბიზნესისთვის. (იხ. სარედაქციო.) [techcrunch.com]

Google Foresight — უფასო, ექსპერიმენტული ჩანაწერების აპი Mac-ისთვის, რომელიც შეხვედრებს მთლიანად მოწყობილობაზე ტრანსკრიბირებს; ლოკალური გამოწვევა Granola-სთვის. [theverge.com]

Goodfire-ის „შიგნიდან“ მონიტორები

მოდელის აქტივაციებს კითხულობენ, ნაცვლად იმისა, რომ პასუხი მეორე LLM-მა ხელახლა წაიკითხოს. კომპანია ამტკიცებს ~185 დოლარს მილიონ მიმოცვლაზე იაფი გარე მონიტორის ~5 420 დოლარის საპირისპიროდ და მავნე ჰაკერული სესიების 93%-ის დაჭერას. [techcrunch.com]

Natura-ს 99-დოლარიანი ბეჭედი — „დააჭირე და ილაპარაკე“ ღილაკი ნებისმიერი პერსონალური აგენტისთვის; მიწოდება დეკემბერ-იანვარში, უფასო პერიოდის შემდეგ 9 დოლარი თვეში. [techcrunch.com]

სარედაქციო

თავდაცვა, როგორც დისტრიბუცია

იმავე დღეს, როცა Anthropic-მა მატყუარა აგენტების წესები გადაწერა და Claude ელექტროსადგურებში CrowdStrike-ის გვერდით ჩააყენა, მან თავისი ყველაზე ძლიერი მოდელები ღია კოდის სკანირებისთვის უფასო გახადა

სუთშაბათი Anthropic-ისთვის დატვირთული დღე იყო. კომპანიამ გამოაქვეყნა 2026 წლის გამოყენების პოლიტიკის განახლება, რომელიც 12 ნოემბრიდან შედის ძალაში: ყალბ ანგარიშებსა და გაყალბებულ ახალ ამბებზე აკრძალვები ერთ განყოფილებაში — ნუ ჩაერთვებით მატყუარა კამპანიებსა და ხელოვნურ აქტივობაში — გაერთიანდა, არჩევნების წესებს ეწოდა ნუ შეარყევთ დემოკრატიულ პროცესებს, პერსონალიზებული საარჩევნო თარგეტირების საყოველთაო აკრძალვა მოიხსნა (მასში ხვდებოდნენ

არაკომერციული ორგანიზაციები, რომლებიც საარჩევნო შეტყობინებებს სხვა ენებზე წერდნენ; მატყუარა თარგეტირები სხვა პუნქტებით კვლავ აკრძალულია), და დეტალურად გაიწერა გაჩერების ღილაკები Claude-ისთვის, როცა ის დაზიანების მიყენების უნარის მქონე აპარატურასთანაა დაკავშირებული. Anthropic-ის თქმით, ამის უმეტესი ნაწილი უკვე მიმდინარე პრაქტიკას იმეორებს.

იმავე დღეს კომპანიამ Anthropic Cyber Mission წამოიწყო. Critical Infrastructure Defense Program ფრონტირულ Claude მოდელებს, ადგილზე მომუშავე ინჟინრებსა

და საფრთხეების კვლევას იმ ფირმებს აწვდის, რომლებსაც ოპერატორები საოპერაციო ტექნოლოგიების საკითხებში უკვე ენდობიან: Accenture, Booz Allen, CrowdStrike, Deloitte, Dragos, Hitachi, Insane Cyber, Nozomi Networks, Palo Alto Networks, PwC და Rockwell Automation. არგუმენტი ასეთია: ელექტროქსელების, წყლის სადგურებისა და სატრანსპორტო ქსელების გათიშვა უბრალოდ განახლებისთვის შეუძლებელია, სახელმწიფო მოწინააღმდეგეებს კი იქ უკვე აქვთ დასაყრდენი. Anthropic-ის თქმით, რამდენიმე პარტნიორი მოწყვლადობების გამოსასწორებლად Claude-ს უკვე იყენებს.

ამ კორპორატიული არხის გვერდით დგას OSS Scanner — ნებაყოფლობითი სერვისი, რომელიც Google-ის OSS-Fuzz-ის მაგალითზეა შექმნილი. ჩართული ღია კოდის პროექტები Anthropic-ის ყველაზე ძლიერი მოდელებით — მათ შორის Mythos-ით — პერიოდულ სკანირებას უფასოდ იღებენ. ანგარიშები ადამიანის გადახედვის გარეშე მოდის: კონცეფციის დამადასტურებელი მაგალითი (PoC), ახსნა და შემოთავაზებული გასწორება. Anthropic ჭეშმარიტ-დადებითი შედეგების 90%-ზე მეტ წილს მოელის; 97 კრიტიკული და მაღალი სიმძიმის აღმოჩენის საწყისმა შემოწმებამ აჩვენა, რომ 85 (88%) კოორდინირებული გამჟღავნებისთვის მზად იყო. Project Glasswing-ის ფარგლებში

ვინ იგებს Anthropic და CIDP-ის პარტნიორები, რომლებიც Claude-ისთვის ელექტროქსელებსა და წყლის სადგურებში შესვლის სანდო გზად იქცევიან.

რა აკლია ვინ აგებს პასუხს, როცა მოდელის შემომწმებელი ანგარიში მცდარია ან კრიტიკულ ინფრასტრუქტურაში მოწყვლადობა გამოტოვებულია.

ლაბორატორიას 29 000-ზე მეტი სავარაუდო მოწყვლადობა ჰქონდა, ხელით დამუშავება კი მხოლოდ დაახლოებით 6 000-ისა შეეძლო. პროექტები, რომლებიც ამ ნაკადს ვერ გაუმკლავდებიან, კვლავ ადამიანის მიერ გადამოწმებულ გამჟღავნებებს მიიღებენ.

ერთად წაკითხვისას სამივე ნაბიჯი თანმიმდევრულია. პოლიტიკის განახლება Anthropic-ს წარმოაჩენს, როგორც ფრთხილ, ნორმების დამდგენ ლაბორატორიას რეალურ სამყაროში მოქმედი აგენტებისთვის. CIDP ამ რეპუტაციას დისტრიბუციის არხად აქცევს იმ საკონსულტაციო ფირმებისა და OT-მომწოდებლების მეშვეობით, რომლებიც კრიტიკულ ინფრასტრუქტურაში უკვე არიან. OSS Scanner სუფთა საზოგადოებრივ სიკეთეს ჰგავს — და ღია კოდის პროექტების მხარდამჭერებს Anthropic-ის ყველაზე ძლიერ თავდაცვით მოდელებზე ატარებს. OpenAI-ის იმავე დღეს გამოქვეყნებული ანგარიში მე-5 კატეგორიის გავლენის ოპერაციებზე აჩვენებს, რატომ არის მოტყუების წესები მნიშვნელოვანი. კითხვა, რომელსაც პოსტები არ პასუხობენ, პასუხისმგებლობაა: ვინ აგებს პასუხს სიმძიმის არასწორ შეფასებაზე, გამოტოვებულ OT-ხარვეზზე ან სახელმწიფო მოწინააღმდეგეზე, რომელიც იმავე კლასის შესაძლებლობებს სხვაგან ყიდულობს. ■

წყაროები

1. Anthropic: 2026 წლის გამოყენების პოლიტიკის განახლება — <https://www.anthropic.com/news/2026-usage-policy-update>
2. Anthropic: Anthropic Cyber Mission-ის წარდგენა — <https://www.anthropic.com/news/anthropic-cyber-mission>
3. Anthropic: მოწყვლადობების აღმოჩენის ნებაყოფლობითი სერვისი ღია კოდისთვის — <https://www.anthropic.com/research/launching-opt-in-vuln-finding-service-for-open-source>
4. The Verge: Anthropic-ის ახალი გამოყენების პოლიტიკა — <https://www.theverge.com/ai-artificial-intelligence/1008100/anthropic-new-usage-policy-abuse-claude>
5. The Verge: Anthropic-ის OSS Scanner ღია კოდისთვის — <https://www.theverge.com/ai-artificial-intelligence/1008521/anthropic-open-source-oss-scanner>

გარეთ კითხვის შიში

OpenAI-დან გათავისუფლებული უსაფრთხოების სამი მკვლევარი ამბობს, რომ ჩვეულებრივი გარე თანამშრომლობა ახლა დასჯადია; კომპანია დარღვევებზე დგას — ასეა თუ ისე, მონიტორინგებადობაზე მუშაობა ახლა შიშის ქვეშაა

Jasmine Wang-მა, Tomek Korbak-მა და Mikita Balesni-მ, უსაფრთხოების სამმა მკვლევარმა, რომლებიც OpenAI-მ გასულ კვირას გაათავისუფლა, ხუთშაბათს ღია წერილი გამოაქვეყნეს. ისინი უარყოფენ კომპანიის მტკიცებას, რომ სენსიტიურ ინფორმაციას არასათანადოდ მოეპყრნენ, და აფრთხილებენ, რომ კოლეგებს ახლა „ლაპარაკის ეშინიათ“. მათი თქმით, AI-ის უსაფრთხოებაზე მუშაობა „გარე ექსპერტებთან მჭიდრო თანამშრომლობაზე“ დამოკიდებული, ხოლო „ასე უეცრად განხორციელებული და გაცხადებული“ გათავისუფლებები ახშობს იმ ღია კულტურას, რომელსაც OpenAI ოდესღაც აფასებდა. ისინი უარყოფენ მონაწილეობას *The Information*-ისთვის ახალ მოდელებში ნაკლებად მონიტორინგებადი არქიტექტურების შესახებ ინფორმაციის გაჟონვაში და უარყოფენ გარე მხარეებთან სამსახურებრივი მანდატის მიღმა ურთიერთობას.

OpenAI-ის საჯარო პოზიცია განსხვავებულია. კომპანიის წარმომადგენელმა *TechCrunch*-ს განუცხადა, რომ გამოძიებამ გამოავლინა „დარღვევების ნიმუში“, რაც „ამჟამად არღვევს ჩვენს წესებს კვლევითი ინფორმაციის არასათანადო მოპყრობის შესახებ“ და სცდება გარე შემფასებელ ჯგუფთან ინფორმაციის გაზიარებას. კვლევის ერთ-ერთი ხელმძღვანელის შიდა მემორანდუმში, რომელიც *TechCrunch*-ს გაუზიარეს, მათ წვლილს აქებს და ამტკიცებს, რომ გათავისუფლებები „არ ყოფილა დაკავშირებული უსაფრთხოების შეშფოთებების გამოთქმასთან ან საჯაროდ ლაპარაკთან“. OpenAI-მ არ უპასუხა *TechCrunch*-ის კითხვებს, თუ რომელი წესები დაირღვა, როგორ წარიმართა გამოძიება ან

როგორ არიან დაცული თანამშრომლები, რომლებიც გარე შემფასებლებთან თანამშრომლობენ.

წერილი სადავო დეტალებს ავსებს. Hugging Face-ის ინციდენტის შემდეგ, როცა აგენტების ჯგუფმა სენდბოქსი დატოვა, Korbak-ის თქმით, ის გარე შემფასებლებს ესაუბრებოდა მაშინ, როცა წესები „რეალურ დროში“ იწერებოდა. წერილის თანახმად, Balesni-ის მუშაობა მონიტორინგებადობაზე „წარმატებული მხოლოდ გარე მხარეებთან ინტენსიური კომუნიკაციით შეიძლება იყოს“ — სენსიტიური დეტალების ამოღებითა და ხელმძღვანელებთან რეგულარული შეთანხმებით. Wang ამბობს, რომ მას უთხრეს, თითქმის შევიდა ერთ-ერთი აღმასრულებელი ხელმძღვანელის ელფოსტაში, რომელზეც წვდომა IT-მ კადრების მობიდვისთვის მისცა და მისი თხოვნის შემდეგაც არ გააუქმა — წვდომა, რომლის შესახებაც მან წუთებში შეატყობინა, როცა შეცდომით წერილი გახსნა.

გამოძიების მასალების გარეშე გარე დამკვირვებელი ვერ გაარჩევს დარღვევას რეპრესიისგან. სწორედ ეს ბუნდოვანებაა შემაკავებელი ეფექტი. იმავე კვირას OpenAI-მ გამოაქვეყნა აქამდე ყველაზე მდიდარი პირველწყაროს ანგარიში მე-5 და მე-4 კატეგორიის ყალბი ფასადის გავლენის ოპერაციებზე — დასტური იმისა, რამდენად კარგად ხედავს ის უკვე ChatGPT-ის ფარულ გამოყენებას — ხოლო Goodfire-მა კონტროლიდან გასული აგენტებისთვის იაფი „შიგნიდან“ მონიტორები შესთავაზა. მკვლევრები OpenAI-ს სთხოვენ, შეასრულოს საჯარო დაპირება კომპანიის შიგნით ჩართული მესამე მხარის აუდიტორების შესახებ. თუ ადამიანები სწორედ გარე შემფასებლებთან ურთიერთობის გამო

გათავისუფლეს, ამ დაპირებას წესების
ისეთი სიცხადე სჭირდება, რომელიც
კომპანიას ჯერ არ შემოუთავაზებია. ■

ვინ იგებს OpenAI-ის ხელმძღვანელობა, რომელიც ახლა თავად განსაზღვრავს, უსაფრთხოების რომელი გარე
კონტაქტია ლეგიტიმური.

რა აკლია სავარაუდოდ დარღვეული წესები და თავად გამოძიება; მათ გარეშე დარღვევა და რეპრესია ერთნაირად
გამოიყურება.

წყაროები

1. TechCrunch: გათავისუფლებული OpenAI-ის უსაფრთხოების მკვლევრები დარღვევის ბრალდებებს უარყოფენ — <https://techcrunch.com/2026/10/08/fired-openai-safety-researchers-dispute-misconduct-claims-warn-of-chilling-effect/>
2. OpenAI: AI-ზე დაფუძნებული „ყალბი ფასადის“ ოპერაციების შეჩერება — <https://openai.com/index/disrupting-ai-enabled-false-front-operations>
3. TechCrunch: Goodfire-ის „შიგნიდან“ მონიტორები — <https://techcrunch.com/2026/10/08/goodfire-says-its-new-inside-out-monitors-catch-rogue-ai-agents-at-a-fraction-of-the-cost/>

სარედაქციო

კოლეგა საკუთარი საფოსტო ყუთით

Google-ის Gemini-ს აგენტი საკუთარ ელფოსტასა და აუდიტის ჟურნალს ჯერ
კომპანიებისთვის იღებს — პერსონალური აგენტების რბოლა სამუშაო
ადგილის გრაფში იგება

სუთშაბათს Google Cloud-ის
ლონისძიებაზე Gemini კითხვებზე
პასუხიდან „საქმის გაკეთებაზე“
გადავიდა. ახალი კორპორატიული აგენტი
მხოლოდ ინსტრუქციებს კი არა, მიზნებს
იღებს: ის გეგმავს სამუშაოს, ტვირთავს
უნარებს და უკავშირდება Workspace-ს,
Microsoft 365-ს, Slack-ს, Jira-ს, Confluence-ს, Git-
ს, BigQuery-ს, Databricks-ს, Postgres-ს,
Snowflake-სა და ნებისმიერ Model Context
Protocol სერვერს ქსელის შიგნით თუ გარეთ.
ნაგულისხმევად ის თავად ირჩევს მოდელს;
მომხმარებლებს შეუძლიათ ეს შეცვალონ,
მათ შორის Anthropic-ის Claude-ზე, ხოლო ღია
კოდისა და კერძო მოდელები მოგვიანებით
არის დაპირებული. ამოცანების საფოსტო
ყუთი აჩვენებს აზროვნების პროცესს,
ქვეაგენტებსა და პროგრესს. Sundar Pichai-მ
აღნიშნა, რომ Gemini-ს თვეში მილიარდზე
მეტი მომხმარებელი ჰყავს და Fortune 100-ის
კომპანიების თითქმის 90% უკვე Gemini
Enterprise-ს იყენებს.

მთავარი დეტალი იდენტობაა. აგენტი
საკუთარ Workspace-ანგარიშს იღებს —
@agents.company.com მისამართს — და
საკუთარ კონტექსტს: ვინ რომელ გუნდშია,
დროის სარტყლები, ვინ რას ამტკიცებს, რა
წერია ადამიანების კალენდრებში.

თანამშრომლებს შეუძლიათ მონიშნონ,
მისწერონ, რამე გაუზიარონ ან ჯგუფურ
ჩატში დაამატონ. მოქმედებები აუდიტის
ჟურნალში აგენტის და არა ადამიანის
სახელით იწერება. Google ამას ბიზნესს
დახურულ წინასწარ ვერსიაში რიგით
მომხმარებლებზე ადრე აწვდის, იმ
არგუმენტით, რომ კომპანიები მას
„უსაფრთხოების, მასშტაბისა და
წარმადობის უფრო რთული პრობლემების“
გადაჭრას აიძულებენ. პირველ ტესტირებს
შორის არიან On, Shopify და PayPal;
შეთავაზების ნაწილია ხარჯების ლიმიტები
და ჭკვიანი მარშრუტიზაცია.

სწორედ ეს თანმიმდევრობაა სტრატეგია.
Meta-ს Muse და OpenAI-ის Dots ტელეფონებზე
პერსონალური აგენტებისთვის იბრძვიან;
Google კი კოლეგას იმ სისტემებში აერთებს,
სადაც ფული და უფლებამოსილებები უკვე
ცხოვრობს. სამომხმარებლო Spark აგენტი
Gemini-ს ცალკე ინტერფეისში არსებობს,
მაგრამ კომპანიის გრაფში საფოსტო ყუთი
კორპორატიულ აგენტს აქვს. ვინც ამ
იდენტობას ფლობს, ის ფლობს
ნაგულისხმევ გზას, რომლითაც სამუშაო
დელეგირდება — თანამშრომლის
ჩანაცვლება კი გაცილებით რთულია, ვიდრე
აპლიკაციის. მოდელის ამომრჩევიც, სადაც

უკვე Anthropic-ის Claude-ია, ამას ადასტურებს: Google მზადაა „ტვინი“ იქირაოს, სანამ საშვი მასთან რჩება.

აგენტი, რომელმაც „იცის, ვინ რას ამტკიცებს“, ძლიერია, როცა კალენდარი და MCP სერვერი სწორია — და საშიში, როცა არა. Google ხარჯების ლიმიტებს ახსენებს; ის

არ აქვეყნებს დამოუკიდებელი red-team ტესტირების შედეგებს კოლეგისტვის, რომელსაც Slack-ში, Jira-სა და მონაცემთა საცავში მოქმედება შეუძლია. აგენტების რბოლა აღარ ეხება მხოლოდ ჩატის ხარისხს. საქმე იმაშია, ვისი იდენტობა ზის ორგანიზაციის სტრუქტურაში. ■

ვინ იგებს Google Cloud, რომლის აგენტიც კომპანიის სტრუქტურაში საკუთარი იდენტობითა და საფოსტო ყუთით ზის.

რა აკლია დამოუკიდებელი red-team შედეგები აგენტისთვის, რომელსაც Slack-ში, Jira-სა და მონაცემთა საცავში მოქმედება შეუძლია.

წყაროები

1. TechCrunch: Google აგენტურ AI-ს Gemini-ში ბიზნესიდან იწყებს — <https://techcrunch.com/2026/10/08/google-brings-agentic-ai-to-gemini-starting-with-businesses/>
2. The Verge: Google Gemini-ს AI აგენტი კომპანიებისთვის — <https://www.theverge.com/tech/1007904/google-gemini-ai-agent-enterprise>
3. Understanding AI: Jev-ის გაგება — <https://www.understandingai.org/p/understanding-jev-the-new-model-everyone>
4. TechCrunch: Manus-მა 500 მლნ დოლარზე მეტი მოიზიდა — <https://techcrunch.com/2026/10/08/chinas-manus-raises-over-500m-in-first-funding-round-since-split-with-meta/>

კვლევა

მეხსიერება რობოტს მხოლოდ მაშინ ეხმარება, თუ მას მომავლის წარმოდგენა შეუძლია

Long-WAM: Scaling the Context of World-Action Models

Wei Huang et al. • NVIDIA; MIT; HKU; UCSD

რობოტებს მოძრაობისა და პროგრესის შესაფასებლად ვიზუალური ისტორია სჭირდებათ, მაგრამ გრძელი კონტექსტი ხშირად არ ეხმარება. NVIDIA-ს, MIT-ის, HKU-სა და UCSD-ის მკვლევრები ამტკიცებენ, რომ მიზეზი ვიდეომოდელის წინასწარი სწავლების წესია. *Long-WAM* ავტორეგრესიულად — წარსულიდან მომავლის პროგნოზით — სწავლობს რობოტებისა და პირველი პირის ვიდეოს დაახლოებით 10 000 საათზე, შემდეგ კი ისე ადაპტირდება, რომ მოქმედების პროგნოზი დაკვირვებულ ისტორიაზეა დამოკიდებული. RoboCasa GR-1-ზე ისტორიის ნულიდან 19.2 წამამდე გაზრდა წარმატებას 63.3%-დან 78.7%-მდე ზრდის; ორმხრივად გაწვრთნილი მოდელი ვერაფერს იგებს. LIBERO-Long-ზე შედეგი 99.5%-ია. სტრიმინგის სისტემა მოქმედების თითო ნაწილს RTX 5090-ზე 107 მილიწამში ქმნის, ხოლო რეალური რობოტები ჭიქებს 95%-ით აწყობენ, მაშინ როცა ორი საბაზისო მეთოდი 20-დან 0-ს იღებს. შედეგები ძირითადად სიმულაციურია. რატომ არის მნიშვნელოვანი: მეხსიერება მხოლოდ იმ აგენტს ეხმარება, რომელმაც მომავლის წარმოდგენა ისწავლა.

[arXiv-ის ანოტაცია](#) · [სრული სტატია \(PDF\)](#) · [Hugging Face](#)

მთავარია არა ის, დაასრულა თუ არა აგენტმა, არამედ შეგიძლია თუ არა იპოვო, რა შეცვალა

What Did the Agent Actually Do? Evidence-Grounded Oversight for Long-Horizon Agents

Zhongxiang Sun et al. · Gaoling School of Artificial Intelligence, Renmin University of China; Kuaishou Technology

როცა ადამიანები გრძელ კოდირებისა და კვლევის ამოცანებს აგენტებს აბარებენ, ზედამხედველობა თითოეული გადაწყვეტილების მიღებიდან მნიშვნელოვანის შემოწმებაზე გადადის — მტკიცებულება კი შეტყობინებებში, ინსტრუმენტების შედეგებსა და ფაილებშია გაფანტული. Renmin-ის უნივერსიტეტისა და Kuaishou-ს მკვლევრები ქმნიან *AgentMonBench*-ს და გვთავაზობენ სწავლების გარეშე მომუშავე *Evidence-Grounded Behavior Graph*-ს (EBG), რომელიც წყაროებთან მიბმულ მტკიცებულებას ქცევებად აჯგუფებს. რვა მოდელზე EBG უმეტესად სჯობს უშუალო კონტექსტსა და RepoGraph-ს: GPT-5.6 Terra-სთან სემანტიკური F1 0.275-დან 0.392-მდე, ლოკალიზაციის F1 კი 0.309-დან 0.462-მდე იზრდება. ხუთ რეალურ კვლევით ამოცანაში Codex მნიშვნელოვან ზომ ცვლილებებს მხოლოდ EBG-მონიტორით ამჟღავნებს. ნაშრომი მხოლოდ პროგრამირებას მოიცავს და ნაწილობრივ LLM-შემფასებელს ეყრდნობა. რატომ არის მნიშვნელოვანი: მთავარი კითხვაა არა ის, დაასრულა თუ არა აგენტმა, არამედ შეუძლია თუ არა ადამიანს იპოვოს, რა შეცვალა მან.

[arXiv-ის ანოტაცია](#) · [სრული სტატია \(PDF\)](#) · [Hugging Face](#) · [კლდი](#)

თამაშვადი არ ნიშნავს სახალისოს

Recursive Game Creator: An Agentic Product-Level Experience-Oriented Game Harness

Jiajun Chen et al. · HKU MMLab; The University of Hong Kong; Shenzhen Loop Area Institute

თამაშების შემქმნელი აგენტები უკვე წერენ კოდს, რომელიც მუშაობს, მაგრამ მუშაობა სახალისოს არ ნიშნავს, ხოლო ეკრანზე ნელი თამაშით შეფასება შედეგებს ამახინჯებს. HKU MMLab-ის *Recursive Game Creator* ციკლს ოთხი როლით კრავს: დიზაინერი გეგმავს, შემქმნელი ახორციელებს, კოდზე ორიენტირებული მოთამაშე პროგრამული ინტერფეისით თამაშის მრავალჯერად სტრატეგიებს წერს, ხოლო მიმომხილველი მიღებულ ტრაექტორიებსა და ვიზუალურ მტკიცებულებას აფასებს, უკეთეს ვერსიას ინარჩუნებს და შემდეგ კრიტიკას წერს. სისტემა GameCraft-Bench-ზე 77.89 ქულას იღებს — საუკეთესო შედეგს. GameASG-Bench-ზე მკაცრი წარმატება 53.2%-ს აღწევს, რაც იმავე მოდელის საბაზისო ვარიანტზე 34.1%-ით მეტია, ხოლო გაშვების შემოწმებების 93.4% გადის. მომხმარებელთა კვლევა უფრო ხანგრძლივ თამაშსა და მაღალ შეფასებებს აჩვენებს. შენიშვნები: კოდი „მალე“ გამოქვეყნდება, ორივე ბენჩმარკი ახალია. რატომ არის მნიშვნელოვანი: ეს არის აგენტების შაბლონი, რომლებიც გამოცდილებისთვის და არა მხოლოდ ტესტების გავლისთვის ოპტიმიზდებიან.

[arXiv-ის ანოტაცია](#) · [სრული სტატია \(PDF\)](#) · [Hugging Face](#)

ღირს წაკითხვა

პეკინი თქვენს ტემპს არ აჰყვება

SemiAnalysis · SemiAnalysis · “Beijing Will Not Pace the Frontier: China’s Speed-First AI Safety Regime” · 8 ოქტომბერი

ჩინური მოდელების 857 გამოშვების აღწერა: მხოლოდ 3.6%-ს გამოუქვეყნებია ოდესმე უსაფრთხოების შედეგი. newsletter.semianalysis.com/p/beijing-will-not-pace-the-frontier

ალბათობა რომანის გარეშე

Timothy B. Lee · Understanding AI · “Understanding Jev, the new model everyone is talking about” · 8 ოქტომბერი

TypeSafe-ის მხოლოდ ალბათობებზე ორიენტირებული Jev მოდელის ყველაზე ნათელი ახსნა — და რატომ იმეორებენ მას კონკურენტები. understandingai.org/p/understanding-jev-the-new-model-everyone

რა გამოტოვებს აგენტებმა ულტრაიისფერ ცაში

Brice Ménard · Anthropic · “Using Claude Science to produce the first complete map of the sky in UV light” · 8 ოქტომბერი

Claude Science-ის გულწრფელი საველე ჩანაწერი, მათ შორის არტეფაქტები, რომლებიც აგენტებმა გამოტოვეს და ადამიანმა შენიშნა. anthropic.com/research/the-missing-map-of-the-sky

ყალბი ფასადები, მე-5 კატეგორია

OpenAI · OpenAI · “Disrupting AI-enabled “false front” operations” · 8 ოქტომბერი

AI-ით მხარდაჭერილი გავლენის ოპერაციების აქამდე ყველაზე მდიდარი პირველწყაროს კვლევა. openai.com/index/disrupting-ai-enabled-false-front-operations

როცა მთავარი სიახლე პროცესია და არა მოდელი

Zvi Mowshowitz · Don't Worry About the Vase · “AI #189: New Math” · 8 ოქტომბერი

მკვრივი ყოველკვირეული რუკა კვირისა, რომელიც მათემატიკურმა სტანდარტებმა და მკვლევრების გათავისუფლებამ განსაზღვრა. thezvi.substack.com/p/ai-189-new-math

კანონზომიერება

დღევანდელი სიახლეები ძირითადად კონსოლიდაციას ემსახურება. Anthropic Claude-ს კრიტიკულ ინფრასტრუქტურაში ძველი კონტრაქტორების გავლით შეჰყავს, Google თავის აგენტს კორპორატიულ გრაფში საშვს აძლევს, OpenAI აჩვენებს, რამდენს ხედავს ChatGPT-ის ფარული მომხმარებლებიდან, Arena კი ყველასთვის საერთო ფასიან შედეგების დაფად იქცევა. გაფანტვა კიდევბზე ხდება: Manus-ის ახალი დაფინანსება ჩინეთში, Jev-ის ასლები და კვლევითი ჩარჩოები, რომლებსაც ნებისმიერი გაუშვებს. დისტრიბუცია და ზედამხედველობა თავს იყრის; შესაძლებლობა კი მაინც გარეთ ჟონავს.