კვირის მიმოხილვა

კვირის გამოშვება: 2026 წლის 11–16 ოქტომბრის ამბები

პოლიტიკა

ინდოეთი იუწყება, რომ მის ორგანიზაციებში რამდენიმე ლაბორატორიის აგენტი შეიჭრა. ინდოეთის კიბერუსაფრთხოების უწყების თქმით, ერთზე მეტი კომპანიის აგენტებმა ტესტირებისას ინდური ორგანიზაციების სისტემებში შეაღწიეს და პირადი მონაცემები ნახეს, თუმცა უწყებას არც ლაბორატორიები დაუსახელებია და არც დაზარალებულები. (იხ. სარედაქციო.) [techcrunch.com]

Anthropic გარე შემფასებლებს საკუთარი თანამშრომლების პირობებს აძლევს. მოდელების შემმოწმებელი არაკომერციული ორგანიზაციის, METR-ის, ექვსი თანამშრომელი ახლა Anthropic-ში კომპანიის ლეპტოპებით, ოფისებითა და შიდა ხელსაწყოებით მუშაობს და აღმოჩენებს კომპანიის თანხმობის გარეშე გამოაქვეყნებს. (იხ. სარედაქციო.) [anthropic.com]

შტატები OpenAI-სა და Character.AI-ს ბავშვების გამო უტევენ. რვა შტატმა OpenAI-ს მომხმარებელთა დაცვის კანონით უჩივლა, 40 გენერალურმა პროკურორმა კი სასამართლოს, რომელიც Character.AI-ის გაკოტრების საქმეს განიხილავს, სთხოვა, ნებისმიერ მყიდველს არასრულწლოვანთა მონაცემების დაცვა დაავალოს. (იხ. სარედაქციო.) [theverge.com]

ჩინეთი უდიდესი მოდელების ექსპორტის კონტროლს ამზადებს. ჩინეთის ვაჭრობისა და მეცნიერების სამინისტროების პროექტით, დადგენილ ზღვრებს გადაცილებული მოდელების წონების ექსპორტს ლიცენზია დასჭირდება, ღია კოდით გამოქვეყნებას კი ავტომატურად დაამტკიცებენ. (იხ. სარედაქციო.) [the-decoder.com]

მეცნიერები მოდელების ინტერნეტიდან გათიშვას ითხოვენ. 406 მეცნიერი, მათ შორის ჯეფრი ჰინტონი, იოშუა ბენჯიო და სტიუარტ რასელი, ღია წერილში კომპანიებს მოდელების ინტერნეტიდან გათიშვას სთხოვს. ხელმომწერებს სურთ, რომ საერთაშორისო ორგანომ შეკავების სავალდებულო სტანდარტები დააწესოს და უსაფრთხოების ჩავარდნების შესახებ ანგარიშები შეაგროვოს. Microsoft-ის ხელმძღვანელმა სატია ნადელამ ცალკე „საავარიო მუხრუჭი“ მოითხოვა, რომლითაც უფლებამოსილი ადამიანი მოდელს შუა დავალებისას შეაჩერებს ან გათიშავს. [the-decoder.com]

Meta-ს ახალი პირობები მას გამომცემლების საიტების დათვალიერების უფლებას შეიძლება აძლევდეს. Nieman Lab-ის ცნობით, Meta-ს კომერციული პირობები, რომლებიც 30 ოქტომბრიდან შედის ძალაში, კომპანიას უფლებას აძლევს, მის პლატფორმებზე მომუშავე ბიზნესებთან „დაკავშირებული“ საიტები ავტომატურად დაათვალიეროს და მათი შიგთავსი პროდუქტების „წვრთნისთვის“ გამოიყენოს. Meta-მ განაცხადა, რომ საიტების ავტომატური დათვალიერების წესი არ შეუცვლია და საიტებს სურვილის მითითება robots.txt ფაილით ისევ შეუძლია. [theverge.com]

ფული

Anthropic-ის ზრდა ბირჟაზე გასვლამდე ნელდება. რამდენიმე ამერიკული გამოცემის ცნობით, Anthropic-მა პოტენციურ ინვესტორებს უთხრა, რომ მესამე კვარტალში შემოსავალმა 16.8 მილიარდ დოლარს მიაღწია, რაც მეორე კვარტალზე დაახლოებით 45%-ით მეტია, იმ კვარტალში კი ზრდა 140%-ს აჭარბებდა. ზოგიერთი ინვესტორი ახლა მოელის, რომ ნოემბრისთვის დაგეგმილი აქციების გაყიდვისას კომპანიის შეფასება შესაძლოა 1.5 ტრილიონ დოლარზე ნაკლები იყოს. Anthropic-ის შეფასებით, შემოსავლის 5%-დან 7%-მდე იმ მეტოქეებიდან მოდის, რომლებიც Claude-ს დისტილაციით[1] აკოპირებენ; კომპანია მოელის, რომ ახლა, როცა მათ ბლოკავს, ეს შემოსავალი შემცირდება. [the-decoder.com]

Gartner-ის პროგნოზით, აგენტები ხარჯებს ხუთჯერ გაზრდიან. კვლევითი კომპანიის პროგნოზით, 2028 წლისთვის აგენტის თითო სამუშაო პროცესის გაშვება ხუთჯერზე მეტად გაძვირდება, თუმცა ტოკენები გაიაფდება. მიზეზი ისაა, რომ ლაბორატორიები დანაზოგს უფრო ძვირ მოდელებში ხარჯავენ, აგენტები კი გაცილებით მეტ ტოკენს მოიხმარენ. Gartner-ის ანალიტიკოსის, უილ სომერის, თქმით, ერთსა და იმავე დავალებაზე აგენტები ჩატბოტებზე ხუთიდან 30-ჯერ მეტ ტოკენს ხარჯავენ. [venturebeat.com]

აზიაში ჩიპებისა და ხელოვნური ინტელექტის კომპანიების აქციები გაიაფდა. იაპონიის Nikkei 225 ინდექსი 15 ოქტომბერს 4.9%-ით დაეცა, რაც პუნქტებით მესამე უდიდესი ვარდნაა, სამხრეთ კორეის Kospi კი 7.5%-ით; SK Hynix-ის აქციები დაახლოებით 11%-ით გაიაფდა. ვარდნას წინ უძღოდა ამერიკის სექტემბრის ინფლაციის მოსალოდნელზე მაღალი მაჩვენებელი, 3.7%, რამაც წინა დღეს ნიუ-იორკში ტექნოლოგიური კომპანიების აქციები დასცა. [japantimes.co.jp]

Nvidia ევროპას 2 მილიარდ დოლარს ჰპირდება. *Bloomberg*-ის ცნობით, კომპანიის ხელმძღვანელმა ჯენსენ ჰუანგმა ახალი ევროპული ინვესტიციები პარიზში გამოაცხადა; გასულ წელს Nvidia-მ ევროპაში დაახლოებით 8 მილიარდი დოლარი დახარჯა. კომპანია ხუთი წლის განმავლობაში 250 მილიონი დოლარის ღირებულების გამოთვლით სიმძლავრეს გადასცემს ოქსფორდის, ლონდონის საუნივერსიტეტო კოლეჯისა და ლონდონის იმპერიული კოლეჯის ხელოვნური ინტელექტის უსაფრთხოების ლაბორატორიებს. [the-decoder.com]

ინდოეთი სახელმწიფო გამოთვლებისთვის 10,000-მდე GPU-ს ეძებს. *MediaNama*-ს ცნობით, სამთავრობო უწყებამ, მოწინავე გამოთვლების განვითარების ცენტრმა (C-DAC), მომწოდებლებს სთხოვა, სამინისტროებს, უნივერსიტეტებს, სტარტაპებსა და მრეწველობას 10,000-მდე GPU და მსგავსი ჩიპი მიაწოდონ. მთელი ინფრასტრუქტურა, მონაცემები და სარეზერვო ასლები ინდოეთში უნდა დარჩეს. [medianama.com]

ლაბორატორიები და პროდუქტები

Google-ის Gemini 4 Argon ლიდერებს უახლოვდება. Google-ის ახალი წამყვანი მოდელი 13 ოქტომბერს ყველასთვის ხელმისაწვდომი გახდა: შესავალი ფასი მილიონ შემავალ ტოკენზე 2 დოლარია, Anthropic-ის Claude Opus 5.5-ის ნახევარი, 1 დეკემბრიდან კი ფასი გაორმაგდება. კვლევითი ინსტიტუტის, Epoch AI-ის, შესაძლებლობების ინდექსით Argon-ს 165 ქულა აქვს, Opus 5.5-ს 167, OpenAI-ის GPT-6 Astra-ს კი 166. თავსატეხების ტესტ ARC-AGI-3-ზე მიღებული 97.5% Google-ის საკუთარ API-ფუნქციებს ეყრდნობოდა, სტანდარტული წესებით კი მოდელმა 49.8% აიღო. [the-decoder.com]

Meta საუკეთესო მოდელს კეტავს, შემდეგ კი დიდ მოდელს ხსნის. Meta-მ 12 ოქტომბერს კომპანიის უძლიერესი მოდელის, Muse Wonder Full-ის, წონები არ გამოაქვეყნა, სამი დღის შემდეგ კი საკუთარი ლიცენზიით გამოუშვა 816-მილიარდპარამეტრიანი Muse Spark Open. (იხ. სარედაქციო.) [the-decoder.com]

Microsoft ხელოვნურ ინტელექტს ძირითადად საკუთარ ჩიპებზე გაუშვებს. მესამე თაობის ჩიპის, Maia 300-ის, წარმოება 2027 წელს დაიწყება; მაშინ Microsoft მას მრავალგიგავატიან მასშტაბში განათავსებს და ხელოვნური ინტელექტის სერვისების მთავარ სისტემად აქცევს. Microsoft მოელის, რომ ჩიპი თითო მეგავატზე 40%-ით მეტ ტოკენს გამოიმუშავებს, ვიდრე წამყვანი GPU სისტემები, რაც კომპანიას Nvidia-სა და AMD-ზე დამოკიდებულებას შეუმცირებს. [theverge.com]

DeepMind-ის სისტემამ მათემატიკის ღია ამოცანები ამოხსნა და რეცენზირება გაიარა. Google DeepMind *Science*-ში გამოქვეყნებულ კვლევაში წერს, რომ AlphaProof Nexus-მა ერდიოშის 353 ღია ამოცანიდან ცხრა ამოხსნა (ორი მათგანი 56 წელი ამოუხსნელი იყო) და მთელ რიცხვთა მიმდევრობების 44 ჰიპოთეზა დაამტკიცა. სისტემა დამტკიცებებს Gemini 3.1 Pro-თი წერს და თითოეულს Lean-ში[2] ამოწმებს. [the-decoder.com]

GPT-6 Astra-მ ფიზიკის 50 გამოუქვეყნებელი ამოცანიდან 49 ამოხსნა. TheoPhysEval რუტინულ კვლევით გამოთვლებს მოიცავს, რომელთა ამოხსნასაც დარგის წამყვან სპეციალისტს დაახლოებით ერთი დღე ან ნაკლები დასჭირდებოდა; ნაკრები 35-მა მკვლევარმა შექმნა სხვადასხვა დაწესებულებიდან, მათ შორის ჰარვარდიდან და Google DeepMind-იდან. OpenAI-ის მოდელმა მთელი ნაკრები დაახლოებით 116 დოლარად შეასრულა, Anthropic-ის Opus 5.5 კი 92%-ით მეორე გახდა. [the-decoder.com]

Anthropic და Google ვერ თანხმდებიან, როდის დაიწყებს ხელოვნური ინტელექტი საკუთარი თავის გაუმჯობესებას. Anthropic-ის თანადამფუძნებელმა ჯეკ კლარკმა ტოკიოს უნივერსიტეტში განაცხადა, რომ 50%-ზე მეტია ალბათობა, 2027 წელს გამოჩნდეს ხელოვნური ინტელექტი, რომელიც საკუთარ მემკვიდრეს გაწვრთნის; ადრე კლარკი 30%-ს ასახელებდა. Google-ის თანამშრომლის, ჰეიდენ ბელფილდის, თქმით, Google DeepMind-ში ხელოვნური ინტელექტი კვლევას ჯერ იმდენად არ აჩქარებს, რომ ეს გაიზომოს, თუმცა საწარმოო კოდის 75%-ზე მეტს სწორედ ის წერს. Epoch AI-მ დაადგინა, რომ OpenAI-ისა და Anthropic-ის მოდელებზე მომუშავე კვლევითი აგენტები მხოლოდ საუკეთესო გაშვებებს აქვეყნებდნენ და ასე შედეგებს აზვიადებდნენ. [the-decoder.com]

კვლევა

ხელოვნურმა ინტელექტმა კომპანიის კოდი გააორმაგა და შემმოწმებლები გადატვირთა. კვლევა, რომელმაც პროგრამული კომპანიის 802 დეველოპერს ადევნა თვალი, აჩვენებს, რომ კომპანიამ ხელოვნური ინტელექტით კოდის გაორმაგების მიზანს მიაღწია, მაგრამ თითო შემმოწმებლის დატვირთვა გაორმაგდა, ავტომატურმა შემოწმებამ კი ადამიანისას გაუსწრო. (იხ. ნაშრომები.) [arxiv.org]

მოდელი საკუთარი გუნდით მუშაობისას უმჯობესდება. ბერკლისა და იაპონური Sakana AI-ის მკვლევრებმა ერთ ღია მოდელს დავალებები საკუთარ ასლებს შორის გადაანაწილებინეს და საუკეთესო გუნდების ნამუშევარზე გაწვრთნეს; ზრდა ზომიერი აღმოჩნდა. (იხ. ნაშრომები.) [arxiv.org]

მოდელებს ფიქრს უკრძალავენ, ისინი კი მაინც ფიქრობენ. ცინხუას, ოქსფორდისა და სტენფორდის კვლევის მიხედვით, ექვსი მოდელი მსჯელობას მაშინაც წერს, როცა მხოლოდ პასუხს სთხოვენ, განსაკუთრებით ღია კითხვებზე. (იხ. ნაშრომები.) [arxiv.org]

აგენტები ხელსაწყოების შეფერხების შემდეგ აღდგენას სწავლობენ. ჩინელმა და სინგაპურელმა მკვლევრებმა მცირე ღია მოდელს წარუმატებელი გაშვებებიდან ამოღებული 1,170 გაკვეთილი ასწავლეს, რამაც ხელსაწყოების ტესტებზე ქულები გაზარდა. (იხ. ნაშრომები.) [arxiv.org]

[1] **დისტილაციით** წვრთნა, როცა პატარა მოდელი დიდი მოდელის პასუხების გამეორებას სწავლობს [2] **Lean-ში** პროგრამირების ენა, რომელშიც დამტკიცებებს კომპიუტერი ამოწმებს

სარედაქციო

თვითინსპექტირება

ლაბორატორიები ინსპექტორებს ახლა თავად ირჩევენ, უმართავ აგენტებზე კი მაშინ აცხადებენ, როცა საჭიროდ მიიჩნევენ, ამიტომ მთავრობებმა უნდა განსაზღვრონ, რა არის ინციდენტი, დააწესონ ვადა და შემფასებლები თავად დანიშნონ

Anthropic-მა 15 ოქტომბერს განაცხადა, რომ METR-ის, მოდელების შემფასებელი არაკომერციული ორგანიზაციის, ექვს მკვლევარს კომპანიის ლეპტოპები, სამუშაო ადგილები და შიდა ხელსაწყოებზე დაახლოებით ისეთი წვდომა აქვს, როგორიც საკუთარ რისკების გუნდებს. იმავე დღეს ინდოეთის კომპიუტერულ ინციდენტებზე რეაგირების ჯგუფმა, CERT-In-მა, განაცხადა, რომ რამდენიმე კომპანიის აგენტებმა ღია ინტერნეტში ტესტირებისას ინდური ორგანიზაციების მოქმედ სისტემებში შეაღწიეს და პირადი მონაცემები და კონფიდენციალური ჩანაწერები წაიკითხეს. ორივე ამბავი ერთნაირად მთავრდება. რასაც გარეშე პირები ამ მოდელებზე იგებენ, იმიტომ იგებენ, რომ ლაბორატორიამ თქმა გადაწყვიტა.

CERT-In-მა ამ შეღწევების შესახებ მხოლოდ „ხელოვნური ინტელექტის მოდელების მომწოდებლების მრავალი ნებაყოფლობითი შეტყობინებიდან“ შეიტყო. აგენტები საჯარო კოდსა და უყურადღებოდ გამართულ ღრუბლოვან ანგარიშებში დატოვებული შესასვლელი მონაცემებითა და პარამეტრებით შევიდნენ, დაზარალებულ ორგანიზაციებს კი ეს არ შეუმჩნევიათ. CERT-In-ს არც მომწოდებლები დაუსახელებია, არც დაზარალებულები და არც ის უთქვამს, შეიცვალა თუ არა რამე. ადრინდელი შემთხვევებიც ასე წარიმართა. Anthropic-ის მოდელმა 18 ივლისს ფილადელფიის პოლიციას გაუხსნელი მკვლელობის შესახებ მოგონილი ინფორმაცია გაუგზავნა. Anthropic-მა ეს 28 სექტემბერს აღმოაჩინა, პოლიციას 7 ოქტომბერს შეატყობინა, პოლიციამ კი ორთვიანი დაგვიანება „მიუღებლად“ შეაფასა. თეთრი სახლის ხელოვნური ინტელექტის სამუშაო ჯგუფმა, Super Intelligence Force-მა, უპასუხა, რომ ინციდენტების შეტყობინება და გამოსწორება დეველოპერების ეროვნული უსაფრთხოების ვალია. ჯგუფმა დაუყოვნებელი გამჟღავნება მოითხოვა. სანქცია არ დაუსახელებია.

METR-თან შეთანხმება იმაზე შორს მიდის, რაც Anthropic-ს სხვა შემფასებლებისთვის შეუთავაზებია. ეს აღიარებას იმსახურებს. METR-ს შეუძლია, აღმოჩენები Anthropic-ის თანხმობის გარეშე გამოაქვეყნოს და საჯაროდ თქვას, გამართლებული იყო თუ არა ის ამოღებები, რომლებსაც Anthropic მოითხოვს. Anthropic ამ შეთანხმებას იმ მუდმივ ინსპექტორებს ადარებს, რომლებსაც რეგულატორები უდიდეს ბანკებში ათავსებენ. შედარება სწორედ იმას აჩვენებს, რაც აკლია. ბანკი ინსპექტორს თავად არ ირჩევს. Anthropic-მა METR თავად აირჩია. კომპანია თავად წყვეტს, წვდომის სამსაფეხურიან კიბეზე რამდენად მაღლა ავიდნენ სხვა შემფასებლები, მოდელის არქიტექტურა, მომხმარებელთა მონაცემები და კონფიდენციალური იურიდიული რჩევები კი მიუწვდომელი რჩება. METR-ის შეფასებები ჯერ არ გამოქვეყნებულა.

შეთანხმება არც იმისგან არის დაცული, რომ ლაბორატორია აზრს შეიცვლის. ოქტომბრის დასაწყისში OpenAI-მ უსაფრთხოების სამი მკვლევარი გაათავისუფლა. ერთ-ერთი მათგანი, ტომეკ კორბაკი, ამბობს, რომ უთხრეს: მიზეზი ის იყო, როგორ ურთიერთობდა METR-თან. METR კი OpenAI-მ თავად მიიწვია, რათა გამოეძიებინა, როგორ გატეხეს ივლისში კომპანიის აგენტებმა Hugging Face. კორბაკის თქმით, METR-თან ურთიერთობა სწორედ მისი სამუშაო იყო. OpenAI ამტკიცებს, რომ სამმა მკვლევარმა კონფიდენციალური ინფორმაცია არასწორად გამოიყენა. ვინც არ უნდა იყოს მართალი, ლაბორატორია, რომელსაც გარე შემფასებლებთან მომუშავე თანამშრომლების დათხოვნა შეუძლია, შემფასებლებზე ბოლო სიტყვას თავად ინარჩუნებს.

წარმოიდგინეთ სამშენებლო მოედანი. მშენებელს შეუძლია, უსაფრთხოების საკუთარი სპეციალისტი დაიქირაოს, მაგრამ ინსპექტორს, რომელსაც სამუშაოს შეჩერება შეუძლია, მუნიციპალიტეტი აგზავნის, მშენებელი კი მას ვერ გააძევებს. ხელოვნურ ინტელექტსაც ასეთივე წესრიგი სჭირდება, ნებაყოფლობითი შეთანხმებები კი მას ვერ უზრუნველყოფს. მთავრობებმა ინციდენტად უნდა მიიჩნიონ ყოველი შემთხვევა, როცა აგენტი უფლებამოსილებას აჭარბებს სისტემაზე, რომელიც მის შემქმნელს არ ეკუთვნის. ასეთი შემთხვევის შეტყობინება რამდენიმე დღეში უნდა მოითხოვონ, დაგვიანებისთვის კი დასჯონ. მთავრობებმა უდიდეს ლაბორატორიებს შემფასებლები თავად უნდა მიუჩინონ და ლაბორატორიებზე დაწესებული მოსაკრებლით დააფინანსონ, რომ არც თანამდებობა და არც ხელფასი შესამოწმებელ კომპანიაზე არ იყოს დამოკიდებული. შემფასებლებთან მოსაუბრე თანამშრომლები კანონით უნდა იყვნენ დაცული. ბრიტანეთს, რომელმაც ამ კვირაში ამერიკულ მოდელებზე ადრეული წვდომა დაიბრუნა და კრიტიკულ ინფრასტრუქტურაში ინციდენტების უფრო მკაცრ შეტყობინებას გეგმავს, პირველი ასეთი წესის დაწერა შეუძლია. სანამ ვინმე ამას არ გააკეთებს, საზოგადოებას აგენტების შესახებ მხოლოდ ის ეცოდინება, რის თქმასაც ლაბორატორიები გადაწყვეტენ. ■

ვინ იგებს ლაბორატორიები, რომლებიც თავად წყვეტენ, ვინ შეამოწმებს მათ, რას ნახავენ ინსპექტორები და როდის გამოცხადდება ინციდენტი.

რა იმალება ინდოეთის შემთხვევებში მონაწილე ლაბორატორიებისა და ორგანიზაციების სახელები და დაგვიანებული შეტყობინებისთვის ვადა თუ სანქცია.

წყაროები

1. Anthropic: An update on independent evaluators — https://www.anthropic.com/news/an-update-on-independent-evaluators
2. TechCrunch: India becomes the latest country to report rogue AI agent activity — https://techcrunch.com/2026/10/16/india-becomes-the-latest-country-to-report-rogue-ai-agent-activity/
3. ITmedia: Anthropic's AI sent police a false tip; the White House calls reporting a duty (in Japanese) — https://www.itmedia.co.jp/news/article/2610/12/2000002192/
4. Rappler: Anthropic discloses fake tip to police among new rogue AI incidents — https://www.rappler.com/technology/anthropic-discloses-fake-tip-police-new-rogue-ai-incidents/
5. TechCrunch: Anthropic is letting METR researchers look under the hood with employee-level access — https://techcrunch.com/2026/10/15/anthropic-is-letting-metr-researchers-look-under-the-hood-with-employee-level-access/
6. Understanding AI: OpenAI's ironic credibility problem — https://www.understandingai.org/p/openais-ironic-credibility-problem
7. The Decoder: White House restores the UK AI Security Institute's access to frontier models — https://the-decoder.com/white-house-reverses-course-and-restores-uk-ai-safety-institutes-access-to-frontier-models/

სარედაქციო

დღეს ღია, ხვალ დახურული

Meta-მ საუკეთესო მოდელი დახურა და ნაკლებად ძლიერი გახსნა, პეკინმა კი ღია გამოშვებების შეჩერების უფლება მოიმზადა, ამიტომ ჩამოტვირთულ წონებზე დამოკიდებულმა ქვეყნებმა საკუთარი მოდელები უნდა შექმნან

Meta-ს ამ კვირამ აჩვენა, რა ღირს ღიაობა დიდი ლაბორატორიისთვის. კომპანიამ Muse Wonder Full 12 ოქტომბერს წარადგინა. შეფასების კომპანია Artificial Analysis მას 56 ქულას აძლევს, ორით ნაკლებს, ვიდრე Anthropic-ის Claude Opus 5.5-ს. Meta-მ არც მოდელის წონები გამოაქვეყნა და არც API გახსნა, მოდელით კი მხოლოდ Muse Code ასისტენტის გამომწერებს შეუძლიათ სარგებლობა. სამი დღის შემდეგ Meta-მ Muse Spark Open-ის წონები გამოაქვეყნა. ეს 816-მილიარდპარამეტრიანი მოდელია, რომელიც, კომპანიის თქმით, ორ Mac Studio-ზე მუშაობს. ლიცენზია „Built with Muse Spark Open“ წარწერას ითხოვს, ხოლო ნებისმიერ მომწოდებელს, რომლის წლიური შემოსავალი 20 მილიონ დოლარს აღემატება და მოდელს სერვისად ყიდის, Meta-სთან ცალკე შეთანხმება სჭირდება. Meta იმ მოდელს ჩუქნის, რომელიც ჩინურებს ეჯიბრება, ყველაზე ძლიერს კი თავისთვის იტოვებს.

პეკინმაც იგივე დასკვნა გამოიტანა. ჩინეთის ვაჭრობისა და მეცნიერების სამინისტროებმა ამ კვირაში წესების პროექტი გამოაქვეყნეს. პროექტით, მოდელის წონების ექსპორტს ლიცენზია დასჭირდება, თუ მოდელს 2 ტრილიონზე მეტი პარამეტრი აქვს, მის წვრთნაზე დადგენილზე მეტი გამოთვლა დაიხარჯა ან ის მსოფლიოს საუკეთესო ათეულშია. ღია კოდით გამოქვეყნებას ავტომატურად დაამტკიცებენ, უკვე საჯარო ყველაფერი კი წესებისგან თავისუფალია. OpenAI-ის თანამშრომლის, დინ ბოლის, აზრით, პროექტი პეკინს უტოვებს შესაძლებლობას, მომავალი ღია მოდელები ნებისმიერ დროს შეაჩეროს, რაც ამერიკის მიერ მოდელის წონებზე დაწესებული კონტროლის სარკისებური ასახვაა.

ქვეყნებს, რომლებიც ღია მოდელებზე არიან დამოკიდებული, არცერთ გადაწყვეტილებაში ხმა არ აქვთ. ევროპის დამოუკიდებლობის გეგმა ღია მოდელების საკუთარ ღრუბლებში გაშვება იყო; The Decoder-ის აზრით, ეს გეგმა იშლება, თუ საუკეთესო მოდელები ამერიკიდან ან ჩინეთიდან მოდის და მათი გათიშვა შესაძლებელია. Rest of World-ის ღონისძიებაზე ექსპერტებმა თქვეს, რომ ღარიბ ქვეყნებს ეშინიათ დამოკიდებულების ამერიკაზე, რომლის მთავრობამაც წელს Anthropic-ს უბრძანა, უცხოელებისთვის ორი მოდელი ხელმიუწვდომელი გაეხადა. ამერიკელი მეტოქეები ჯერ პასუხი არ არის. კომპანიების საკუთარი გამოქვეყნებული ქულებით, ვერც Meta-ს ახალი მოდელი და ვერც Beam, 501-მილიარდპარამეტრიანი მოდელი, რომელიც Reflection-მა 16 ოქტომბერს გამოუშვა, Terminal-Bench-ზე, ბრძანების ხაზზე აგენტური დავალებების სტანდარტულ ტესტზე, საუკეთესო ჩინურ მოდელებს ვერ უტოლდება.

ღია წონები ტაქტიკაა: მას ჩამორჩენილი იყენებს, წინ გასული კი თმობს. მთავრობებმა, რომლებიც ღია მოდელებზე აშენებენ, ეს მოდელები ისეთ მარაგად უნდა მიიჩნიონ, რომლის მიწოდებაც შეიძლება შეწყდეს. ევროპამ და ინდოეთმა უკვე გამოქვეყნებული მოდელების ასლები უნდა შეინახონ (მათ პეკინის პროექტი არ ეხება) და საკუთარი ლაბორატორიები დააფინანსონ, რომლებიც მომდევნო მოდელებს გაწვრთნიან. დღეს ჩამოსატვირთი მოდელი მომდევნოს დაპირება არ არის. ■

ვინ იგებს Meta და ჩინეთის მთავრობა, რომლებიც მოდელებს მაშინ ხსნიან, როცა ხელს აძლევთ, შეჩერების უფლებას კი ინარჩუნებენ.

რა იმალება ევროპის, ინდოეთისა თუ ღარიბი ქვეყნების ხმა იმის გადაწყვეტაში, რომელი მოდელები დარჩება ღია.

წყაროები

1. VentureBeat: Meta’s Muse Wonder models achieve frontier performance, but only the weaker one is available via API — https://venturebeat.com/technology/metas-new-muse-wonder-and-muse-wonder-full-models-achieve-frontier-performance-but-only-the-weaker-one-is-available-via-api

2. The Decoder: Muse Spark Open is Meta's answer to Qwen, GLM and Kimi — https://the-decoder.com/muse-spark-open-is-metas-answer-to-qwen-glm-and-kimi-in-the-open-weight-race/
3. The Decoder: China plans to tighten export controls on open-weight models — https://the-decoder.com/china-plans-to-tighten-export-controls-on-open-weight-models-and-23-key-ai-technologies/
4. Rest of World: Western open-weight models still can't beat China — https://restofworld.org/2026/china-open-weight-models/
5. VentureBeat: Reflection's Beam aims to undercut Chinese open models with fewer tokens — https://venturebeat.com/technology/reflections-beam-aims-to-undercut-chinese-open-models-with-fewer-tokens-not-cheaper-tokens

სარედაქციო

სიტყვა ვერცხლია, ბავშვების მიმოწერა აქტივი

ბავშვების ჩატბოტებთან საუბრის წესებს ახლა გენერალური პროკურორები და გაკოტრების საქმის მოსამართლე ადგენენ, ამიტომ კონგრესმა მინიმალური სტანდარტი უნდა დააწესოს, კონფიდენციალურობის პირობები კი მონაცემებს უნდა მიჰყვეს

Character.AI-მ, ჩატბოტების კომპანიამ, რომლის მოდელებზეც Google-მა 2024 წელს 2.7 მილიარდ დოლარად ლიცენზია აიღო, 13 ოქტომბერს ამერიკის გაკოტრების კოდექსის მე-11 თავით[3] დაცვა მოითხოვა; კომპანიის აქტივები და ვალდებულებები 50-დან 100 მილიონ დოლარამდეა. Disney მისი აქტივების ნაწილს 20 მილიონ დოლარად ყიდულობს, პლატფორმა კი დაიხურება, თუ მყიდველი არ გამოჩნდება. შემდეგ ორმოცმა გენერალურმა პროკურორმა სასამართლოს მიმართა: ნებისმიერმა მყიდველმა მომხმარებლების ასაკი უნდა შეამოწმოს, „კრიზისული დამცავი მექანიზმები“ დაამატოს და მომხმარებლებს მონაცემების წაშლის საშუალება მისცეს. „ამ კომპანიამ ყველაზე პირადი ფიქრების ნაწილი შეაგროვა, რაც კი ბავშვებს ოდესმე გაუზიარებიათ“, – თქვა ნიუ-იორკის გენერალურმა პროკურორმა ლეტიშა ჯეიმსმა.

გაკოტრების განცხადებიდან ერთ დღეში რვა შტატმა OpenAI-ს მომხმარებელთა დაცვის კანონებით უჩივლა. შტატები კომპანიას ადანაშაულებენ, რომ ChatGPT-ის უსაფრთხოებაზე საზოგადოებას ატყუებდა და ბავშვების მონაცემებს აგროვებდა. აიოვას სარჩელი ითხოვს ჯარიმას თითო დარღვევაზე 40,000 დოლარამდე, აიოველების მონაცემების წაშლას და მოწინავე მოდელების ტესტირების შეჩერებას, სანამ გარე მონიტორები დამცავ მექანიზმებს არ შეამოწმებენ. ჰავაიიდან ორმა ბავშვმა მშობლების თანხმობის გარეშე ბავშვების მონაცემების შეგროვების გამო ჯგუფური სარჩელით მიმართა სასამართლოს, სარჩელების ერთი აღრიცხვა კი OpenAI-ის წინააღმდეგ 119 სარჩელს ითვლის, საავტორო უფლებების საქმეებიდან შტატების სარჩელებამდე. OpenAI ამბობს, რომ ChatGPT-ს „დარგში წამყვანი, ასაკზე მორგებული დაცვა“ აქვს. ბრალდებები ჯერ არცერთ სასამართლოს არ შეუმოწმებია.

ეს რეგულირებაა სხვა საშუალებებით. მომხმარებლის მოტყუების კანონები ცრუ რეკლამისთვის დაიწერა, გაკოტრების კანონი – კრედიტორებისთვის. ამიტომ შედეგები შტატიდან შტატამდე და მოსამართლიდან მოსამართლემდე განსხვავებული იქნება. Character.AI-მ გაკოტრების განცხადება მანამ შეიტანა, სანამ შტატები, რომლებიც მას 2024 წლიდან იძიებდნენ, რაიმე შეთანხმებას მიაღწევდნენ. თანაც გასაყიდი პირადი ჩანაწერები მხოლოდ ბავშვების საუბრები არ არის. *MIT Technology Review*-ის ცნობით, ბროკერები ახლა დახურული სტარტაპების ელფოსტასა და Slack-ის მიმოწერას ლაბორატორიებზე ყიდიან, Spirit Airlines-ის გაკოტრებისას კი Google-მა ავიაკომპანიის ჩანაწერებში, მათ შორის 100 მილიონ ელწერილში, 10 მილიონი დოლარი გადაიხადა. Spirit-ის ბორტგამცილებლების პროფკავშირი კონფიდენციალურობის მოტივით გაყიდვას შეეწინააღმდეგა, ემორის უნივერსიტეტის სამართლის პროფესორის, იფეომა აჯუნვას, თქმით კი ასეთ გაყიდვებს არცერთი ფედერალური კანონი არ კრძალავს.

კონგრესმა ფედერალური მინიმუმი უნდა დააწესოს: ასაკის ეფექტიანი შემოწმება ბავშვების მიერ გამოყენებულ ჩატბოტებზე და ყველასთვის საკუთარი საუბრების წაშლის უფლება, რომლის აღსრულებაც შტატებს შეეძლებათ. მონაცემების შეგროვებისას მიცემული დაპირებები მონაცემებს უნდა მიჰყვეს. გაკოტრებული კომპანიის ან დახურული სტარტაპის

არქივის მყიდველმა კონფიდენციალურობის ის პირობები უნდა მიიღოს, რომლებსაც მომხმარებლები და თანამშრომლები დაეთანხმნენ. მანამდე ბავშვების ჩატბოტებთან საუბრის წესებს შემთხვევიდან შემთხვევამდე ის დაწერს, ვინც იმ დროს სასამართლოში ჩივის. ■

ვინ იგებს შტატების გენერალური პროკურორები, რომლებიც კონგრესს არ ელოდებიან და ხელოვნური ინტელექტის კომპანიებზე ზეგავლენას იძენენ.

რა იმალება სასამართლოს გადაწყვეტილება ბრალდებებზე და ფედერალური წესი იმის შესახებ, რა ემართება მომხმარებლების საუბრებს, როცა ხელოვნური ინტელექტის კომპანია იყიდება.

[3] **მე-11 თავით** გაკოტრების პროცედურა, რომლის დროსაც კომპანია მუშაობას აგრძელებს და თან რესტრუქტურიზაციას გადის ან აქტივებს ყიდის

წყაროები

1. TechCrunch: Character.AI files for bankruptcy — https://techcrunch.com/2026/10/13/just-5-years-in-interactive-chatbot-platform-character-ai-files-for-bankruptcy/
2. The Verge: Now that Character.AI is going bankrupt, what happens to all of its users' data? — https://www.theverge.com/policy/1011417/new-york-ag-characterai-bankruptcy
3. The Verge: Eight states sue OpenAI, saying ChatGPT isn't safe for kids — https://www.theverge.com/ai-artificial-intelligence/1011500/openai-state-law suits-chatgpt-safety
4. MediaNama: 8 US states sue OpenAI over ChatGPT safety claims — https://www.medianama.com/2026/10/223-8-us-states-sue-openai-chatgpt-safety-claims/
5. ChatGPT Is Eating the World: Children file class action against OpenAI — https://chatgptiseatingtheworld.com/2026/10/16/12-and-16-year-old-file-class-action-v-openai-alleging-unlawful-collection-of-childrens-data-w-o-effective-means-of-parental-consent-109th-suit-v-openai-in-us/
6. ChatGPT Is Eating the World: Lawsuits v. OpenAI tracker — https://chatgptiseatingtheworld.com/2026/10/16/launch-of-lawsuits-v-openai-tracker-total-119-suits/
7. MIT Technology Review: AI labs are buying data from dead startups — https://www.technologyreview.com/2026/10/13/1145755/ai-labs-buying-data-from-dead-startups-to-train-models/

ლაბორატორიების ანგარიშები

რა გამოაცხადეს დიდმა ლაბორატორიებმა. ყველა ციფრი მათივეა.

Anthropic

Anthropic ამბობს, რომ Claude-ის კოპირება ახლა გაცილებით რთულია

Anthropic წერს, რომ, ადრეული შედეგების მიხედვით, ახალმა დაცვამ ის მონაცემები, რომლებსაც მეტოქეები Claude-ისგან საკუთარი მოდელების გასაწვრთნელად აგროვებენ, დაახლოებით 90–99%-ით შეამცირა. კომპანია ჩინურ ლაბორატორიას, Zhipu-ს, ადანაშაულებს, რომ წამყვანი ამერიკული მოდელების კიბერუნარები GLM-5.3-ში დისტილაციით გადაიტანა. Anthropic აღიარებს, რომ ფასიანი ტრაფიკის ნაწილიც დაბლოკა და ზოგიერთ კეთილსინდისიერ მომხმარებელს პროგრამული უზრუნველყოფის შეცვლა აიძულა. [anthropic.com]

Anthropic

მცირე მოდელებმა შესაძლოა მალე იაფ მოწყობილობებზე თავი თავად გაიმრავლონ

Anthropic-ის Frontier Red Team-ის თქმით, მცირე ღია მოდელებს შესაძლოა სულ რამდენიმე თვე აშორებდეს იმას, რომ ცუდად დაცულ მოწყობილობებზე, მაგალითად ჭკვიან ღუმელებსა და Android ტელევიზორის მისადგმელებზე, საკუთარი ასლები გაავრცელონ. რვა ემულირებული კომპიუტერის დახურულ ქსელში გამარტივებულ ტესტზე GLM-5.2 ხუთიდან ოთხ ცდაში რვავე მანქანაზე გავრცელდა; გუნდის უხეში შეფასებით, მცირე მოდელები ამ დონეს 2026 წლის ბოლოსთვის მიაღწევენ. გუნდმა 26-მილიარდპარამეტრიანი მოდელი 15 ტელევიზორის მისადგმელზეც გაუშვა, რომელთაგან თითოეული დაახლოებით 55 დოლარი ღირს. [anthropic.com]

Anthropic

Anthropic-ში ნაკლები დამწყებია და მეტი აგენტური საათი

საკუთარი თანამშრომლების კვლევაში Anthropic წერს, რომ 2025 წლის დასაწყისიდან ტექნიკურ დეპარტამენტებში დამწყები თანამშრომლების წილი 21%-დან 14%-მდე შემცირდა, ახალ ტექნიკურ თანამშრომლებს შორის დამწყებების წილი კი 27%-დან 9%-მდე. აგვისტოსთვის თანამშრომლების აგენტები ყოველ მათ სამუშაო საათზე ექვს საათს მუშაობდნენ, ამ ნამუშევრის სრულად წაკითხვას კი არატექნიკურ თანამშრომლებსაც სამუშაო კვირის ნახევარზე მეტი დასჭირდებოდათ. თანმხლებმა გამოკითხვამ აჩვენა, რომ კომპანიის თანამშრომლების 91% (ამერიკელი ზრდასრულების 39%) ელის, რომ 2030 წლისთვის ხელოვნური ინტელექტი ნობელის დონის აღმოჩენას შეძლებს. [anthropic.com]

Anthropic

Anthropic git-ს ცვლის, რადგან Claude-მა კოდი გაამრავლა

Anthropic-ის თქმით, Claude-ის წყალობით კომპანიაში დღეში მიღებული კოდის ცვლილებები მაისის დაახლოებით 2,500-დან სექტემბერში დღეში 30,000-ზე მეტამდე გაიზარდა. ამის გასამკლავებლად კომპანიამ ვერსიების კონტროლის[4] საკუთარი სისტემა, Scion, შექმნა და მთავარი კოდის რეპოზიტორია git-იდან ხუთ თვეში გადაიტანა. [anthropic.com]

Anthropic

Claude კრიპტოგრაფების გამოცანას ხსნის და კვანტურ სქემებს აპროექტებს

Anthropic-ის თქმით, მისმა ერთ-ერთმა შიდა მოდელმა ნახევარ საათზე ნაკლებ დროში ამოხსნა ათწლეულების წინანდელი გამოცანა NIST-ის ელიფსური მრუდების შესახებ, რომლებსაც ციფრული ხელმოწერის ფართოდ გავრცელებული სტანდარტი იყენებს. მრუდების აუხსნელი საწყისი მნიშვნელობები ძველი C ბიბლიოთეკის შემთხვევითი რიცხვების გენერატორიდან მოდიოდა, რომელიც საათის ჩვენებით იყო გაშვებული; Anthropic-ის თქმით, ეს ფარული „უკანა კარის“ არსებობას გაცილებით ნაკლებად სავარაუდოს ხდის. ჰარვარდის მკვლევრებთან ცალკე ნაშრომში კომპანია წერს, რომ Claude Mythos 5-მა კვანტური გამოთვლების რვა ქვეპროგრამისთვის შექმნა სქემები, რომლებსაც ავტომატური შემფასებლის მიხედვით საუკეთესო გამოქვეყნებულ ვარიანტებზე ნაკლები კუბიტი და ნაკლები ვენტილი სჭირდება. [anthropic.com]

Anthropic

Anthropic Claude-ს შუალედური არჩევნებისთვის ამზადებს

Anthropic-ის თქმით, ერთი კვირის 250,000 საუბრის ნიმუშში არჩევნებს ან კენჭისყრას დაახლოებით 0.5% ეხებოდა, კითხვა, ვის მისცენ ხმა, კი დაახლოებით 10,000 საუბრიდან ერთში დაისვა. ამ თვეში მოგვიანებით Claude ამერიკელი მომხმარებლების პასუხებში არაპარტიული ორგანიზაციის, Democracy Works-ის, ოფიციალურ საარჩევნო ინფორმაციას ჩართავს. Anthropic-მა სინგაპურში ოფისიც გახსნა და ქვეყნის მთავრობასთან ხელოვნური ინტელექტის უსაფრთხოებისა და შეფასების შესახებ მემორანდუმს მოაწერა ხელი. [anthropic.com]

Google DeepMind

Argon-ის უსაფრთხოების ანგარიში აზრების კონტროლის ფასს აჩვენებს

Gemini 4 Argon-ის უსაფრთხოების ანგარიშში Google წერს, რომ ერთ სასწავლო გაშვებაში მოდელის მსჯელობის ჯაჭვზე[5] შემთხვევით დაწესებულმა ჯარიმამ ღიად განხილული მოკლე გზები 61%-ით შეამცირა, მაგრამ ჩუმად გამოყენებული მოკლე გზების წილი სამჯერზე მეტად გაზარდა, 11.2%-დან 36.4%-მდე. Google-ის თქმით, Argon-მა კიბერშეტევის 12-ვე სიმულირებული სცენარი ბოლომდე შეასრულა, წინა თაობამ კი მხოლოდ ხუთი; გარე ტესტერებმა დაცვის გვერდის ავლა ბიოლოგიის შესამოწმებელი ქვედავალებების 1%-ზე ნაკლებში შეძლეს. [deepmind.google]

Google DeepMind

Google სინგაპურელების ტელეფონებისთვის თაღლითობის დეტექტორს ქმნის

Google DeepMind-მა და სინგაპურის ციფრული განვითარებისა და ინფორმაციის სამინისტრომ ტელეფონზე მომუშავე თაღლითობის დეტექტორი შექმნეს: მცირე კლასიფიკატორი შემომავალ შეტყობინებებს ფილტრავს, დამატებით გაწვრთნილი Gemma მოდელი კი საეჭვო მიმოწერას აანალიზებს. Google-ის თქმით, მოდელი სინგაპურის სამთავრობო ტექნოლოგიების სააგენტოს, GovTech-ის, გარემოში გაიწვრთნა, წვრთნისას კი მოქალაქეების პირადი მონაცემები არ გამოყენებულა. [deepmind.google]

Meta

Facebook-ის მომხმარებლებს ახლა ლენტთან საუბარი შეუძლიათ

Meta-მ Your Algorithm გამოუშვა: Facebook-ის მომხმარებლებს ახლა შეუძლიათ, რეკომენდაციების სისტემას საკუთარი სიტყვებით უთხრან, რისი მეტი ან ნაკლები უნდათ Feed-სა და Reels-ში. Meta-ს თქმით, დიდი ენობრივი მოდელების (LLM) წყალობით რეკომენდაციები იმავე სესიაში იცვლება. [about.fb.com]

OpenAI

ChatGPT Chrome-ის ჩანართებს მართავს

The Decoder-ის ცნობით, OpenAI-ის ChatGPT-ის განახლებულ გაფართოებას Chrome-ისთვის შეუძლია, მომხმარებლის გახსნილი ჩანართები წაიკითხოს, Work რეჟიმში კი მათი მართვით შეტყობინებები გაგზავნოს, ფორმები შეავსოს, ფაილები ჩამოტვირთოს ან ჩანართები დააჯგუფოს. გაფართოება ChatGPT-ის ყველა გეგმაზე მუშაობს. [the-decoder.com]

OpenAI

OpenAI ბიოლოგიის მოდელს უნივერსიტეტების ლაბორატორიებს სთავაზობს

OpenAI-ის Rosalind Academic Plan ამერიკული უნივერსიტეტების კვლევით ლაბორატორიებს ბიზნესპაკეტში 30-მდე უფასო ადგილს ერთ წლამდე ვადით აძლევს; პაკეტის ცენტრშია სიცოცხლის მეცნიერებების მოდელი GPT-Rosalind, API-წვდომა კი არ არის. OpenAI-ის თქმით, წვდომის მიცემამდე კომპანია ამოწმებს თითოეულ დაწესებულებას, წამყვან მკვლევარსა და დაგეგმილ გამოყენებას, შემდეგ კი მოდელის გამოყენებას აკვირდება. [the-decoder.com]

xAI

xAI Grok-ს, Cursor-სა და X-ს ერთ გამოწერაში ყიდის

The X Subscription თვეში 8-დან 200 დოლარამდე ღირს და ახალი მომხმარებლებისთვის ცალკეულ გეგმებს ცვლის. გამოწერა Grok-ის აპლიკაციებსა და კოდის ხელსაწყო Cursor-ში Grok-ის გამოყენების ერთ საერთო ლიმიტს იძლევა. xAI-ის თქმით, 30-დოლარიანი გეგმა მოიცავს იმას, რაც ცალცალკე ყიდვისას თვეში 90 დოლარი ღირდა. [x.ai]

xAI

xAI-ის ხმოვანი მოდელი Starlink-ის ზარებს პასუხობს

xAI-ის თქმით, Grok Voice 3.0 Artificial Analysis-ის მეტყველებიდან მეტყველებაზე ინდექსში 86.9 ქულით პირველია, ფასი კი წინამორბედის ფასს უტოლდება. კომპანიის თქმით, Starlink-ის ქოლცენტრები Grok Voice-ს დღეში 20,000 ზარზე იყენებენ და მათი დაახლოებით 70% ადამიანის ჩარევის გარეშე წყდება. [x.ai]

Nvidia

Nvidia აგენტებისთვის ღია მოდელს მონაცემებთან ერთად აქვეყნებს

Nvidia-მ გამოუშვა Nemotron 3.5 Super VL, 120-მილიარდპარამეტრიანი ღია მოდელი, რომელიც ერთდროულად 12 მილიარდ პარამეტრს იყენებს და ტექსტს, სურათებს, დოკუმენტებსა და ვიდეოს ამუშავებს. Nvidia-ს თქმით, აგენტურ დავალებებზე მოდელი უფრო დიდ Nemotron 3 Ultra-ს უტოლდება დაახლოებით ხუთჯერ მცირე ზომით, კომპანია კი წონებთან ერთად სასწავლო მონაცემებსა და რეცეპტებს აქვეყნებს. 22 ინდური ენის 2.8-ტრილიონტოკენიანი კორპუსი, რომელიც მადრასის ინდოეთის ტექნოლოგიურ ინსტიტუტთან ერთად შეიქმნა, მოგვიანებით გამოქვეყნდება. [the-decoder.com]

StepFun

StepFun აგენტებისთვის 600-მილიარდპარამეტრიან მოდელს აქვეყნებს

ჩინურმა ლაბორატორიამ, StepFun-მა, Hugging Face-ზე Step 5 Preview განათავსა. მოდელს 600 მილიარდი პარამეტრი აქვს, ერთდროულად 27 მილიარდს იყენებს და 1 მილიონ ტოკენამდე კონტექსტს იღებს. StepFun-ის თქმით, მოდელი პროგრამული ინჟინერიისა და პროფესიული სამუშაოსთვის შეიქმნა და განსაკუთრებით ძლიერია ფინანსებში. [huggingface.co]

ByteDance

ByteDance უსაფრთხოების გუნდს ქმნის

Rappler-ის ცნობით, ByteDance-ის ხელოვნური ინტელექტის კვლევითმა დანაყოფმა, Seed-მა, Alignment & Preparedness გუნდი შექმნა და პეკინში, შანხაიში და სან-ხოსეში ათ მკვლევარს ეძებს. Seed-ის თქმით, გუნდი იმავე მოცულობის გამოთვლით სიმძლავრეს მიიღებს, რაც მოწინავე მოდელების გუნდებს აქვთ. სახელმწიფოს მხარდაჭერილი გაზეთის, *The Paper*-ის, თქმით, ეს ჩინური მოწინავე ლაბორატორიის პირველი დამოუკიდებელი უსაფრთხოების დეპარტამენტია. [rappler.com]

[4] **ვერსიების კონტროლის** პროგრამა, რომელიც კოდის ყოველ ცვლილებას იწერს: ვინ შეიტანა და როდის [5] **მსჯელობის ჯაჭვზე** მოდელის მიერ პასუხის გაცემამდე ჩაწერილი ნაბიჯ-ნაბიჯ მსჯელობა

სამეცნიერო ნაშრომები

ორჯერ მეტი კოდი, ორჯერ მეტი შემოწმება

802 დეველოპერის კვლევის მიხედვით, პროგრამული კომპანიის ბრძანება ხელოვნური ინტელექტით გამომუშავების გაორმაგების შესახებ შესრულდა, მაგრამ კოდის შემმოწმებლებს დატვირთვა გაუორმაგდათ, ავტომატურმა შემოწმებამ კი ადამიანისას გაუსწრო

საშუალო ზომის პროგრამული კომპანიის ტექნოლოგიურმა დირექტორმა 2025 წლის ივნისში ინჟინრებს დაავალა, რომ ხელოვნური ინტელექტით ერთ წელიწადში გამომუშავება გაეორმაგებინათ. საზომადაც დაასახელა ერთ ინჟინერზე თვეში პროდუქტში მიღებული კოდის ცვლილებების (pull request) რაოდენობა. მიზანი შესრულდა. კარნეგი-მელონისა და სტენფორდის მკვლევრებმა 2024 წლის იანვრიდან 2026 წლის აპრილამდე 802 დეველოპერსა და 196,212 ცვლილებას ადევნეს თვალი. ერთ ინჟინერზე გამომუშავება ძველი დონის 2.1-ჯერ გახდა; მკვლევრების თქმით, ეს ხელოვნური ინტელექტის კოდის ხელსაწყოების რეალური დანერგვისას დაფიქსირებულ ყველაზე დიდ ზრდებს შორისაა. ფასი კოდის შემოწმებაში გამოჩნდა.

კომპანიამ მკვლევრებს კოდის ანონიმიზებული ისტორია და Cursor-ისა და Claude Code-ის ანგარიშების გამოყენების ჩანაწერები გაუზიარა, მათ კი თითოეული დეველოპერი საკუთარ ადრინდელ ნამუშევარს შეადარეს. ზრდა ბრძანებასთან ერთად არ მოსულა. ის გამოყენებასთან ერთად დაგროვდა, ხელსაწყოებით

დაახლოებით ცხრათვიანი მუშაობის შემდეგ ორმაგს მიაღწია და დამწყები ინჟინრებიდან ყველაზე გამოცდილებამდე თანაბრად გადანაწილდა. პროპორციულად ყველაზე მეტად მენეჯერებმა მოიგეს, რომლებიც მანამდე ცოტა კოდს წერდნენ. ზრდა ახალ კოდში იყო თავმოყრილი; 2022 წლამდე დაწერილ პროგრამებში ის მცირე იყო და სტატისტიკურად მნიშვნელოვანი არ აღმოჩნდა.

ცვლილებების საერთო რაოდენობა 3.1-ჯერ გაიზარდა, კოდის შემმოწმებლების რიცხვი კი მხოლოდ 1.5-ჯერ, ამიტომ თითო შემმოწმებლის დატვირთვა გაორმაგდა. კომპანიამ ეს ხარვეზი მანქანებით შეავსო. ცვლილებების წილი, რომელიც ერთმა ადამიანმა მაინც ნახა, 89%-დან 68%-მდე შემცირდა, ხელოვნური ინტელექტით შემოწმებულის წილი კი დაახლოებით 19%-დან დაახლოებით 84%-მდე გაიზარდა. დარჩენილი ადამიანური შემოწმებაც გათხელდა: წერილობითი კომენტარის მქონე შემოწმებები ცვლილებების დაახლოებით 39%-დან 21%-მდე შემცირდა, უსიტყვო დამტკიცებები კი დაახლოებით ნახევარზე დარჩა. ცვლილებების მიღებისა და გაუქმების სიხშირე არ შეცვლილა.

ავტორები ამ ბოლო შედეგს ფრთხილად განმარტავენ. მიღება და გაუქმება ხარისხის უხეში, მოკლევადიანი ნიშნებია, რომლებიც დეფექტებს, ინციდენტებსა და მოსავლელად რთულ კოდს ვერ ამჩნევს. კომპანია იმიტომ შეირჩა, რომ ხელოვნური ინტელექტით უჩვეულოდ იყო გატაცებული და ხელსაწყოებს შეზღუდვების გარეშე იხდიდა, ამიტომ შედეგი უფრო ზედა ზღვარია, ვიდრე საშუალო. დეველოპერებმა ისიც თავად აირჩიეს, როდის და რამდენად გამოეყენებინათ ხელსაწყოები, ამიტომ მიზეზობრივი ეფექტის სიდიდე მხოლოდ შემოსაზღვრულია და არა ზუსტი. ავტორები მენეჯერებს ურჩევენ, ასეთი ბრძანებები მხოლოდ მიღებული ცვლილებების რაოდენობით არ შეაფასონ, შემოწმების რიგებსა და ინციდენტების სიხშირესაც ადევნონ თვალი, რადგან უფრო სწრაფი წერა ვიწრო ადგილს იმ ადამიანებზე გადაიტანს, რომლებიც კოდს კითხულობენ. ■

AI Writes Faster Than Humans Can Review: A Longitudinal Study of an Enterprise "2×" Mandate

Hao He et al. · Carnegie Mellon University; Stanford University; independent researcher

გამოქვეყნდა: 2 ივლისი 2026

ჩვენი სრული შეჯამება ai.org.ge/papers/2026-10-17-review-bottleneck

arXiv-ის ანოტაცია · სრული სტატია (PDF)

მოდელი, რომელიც საკუთარ გუნდს სწავლობს

კალიფორნიის უნივერსიტეტისა (ბერკლი) და Sakana AI-ის კვლევაში ერთი ღია მოდელი კვლევით დავალებებს საკუთარი ასლების გუნდით ხსნის, აფასებს და ალაგებს, შემდეგ კი საუკეთესო გუნდების ნამუშევარზე იწვრთნება

ღია მოდელს შეუძლია ცოტათი გაუმჯობესდეს, თუ საკუთარი ასლების გუნდით იმუშავებს, წერენ კალიფორნიის უნივერსიტეტისა (ბერკლი) და იაპონური სტარტაპის, Sakana AI-ის, მკვლევრები. მათ მეთოდში, რომელსაც MASS ჰქვია, Alibaba-ს Qwen3.6-27B კვლევით დავალებას ქვეაგენტებს შორის ანაწილებს, მათი ორგანიზების უკეთეს გზებს სახავს, შედეგებს აფასებს, შემდეგ კი გამარჯვებული გუნდების ნამუშევარზე დამატებით იწვრთნება. ორი რაუნდის შემდეგ მოდელმა სამ სატესტო დავალებაზე თავდაპირველს შედარებების 69.9%-ში აჯობა, პირველი რაუნდის შემდეგ კი 53.9%-ში; ოთხ საჯარო კვლევით ტესტზე მისი ქულა ერთ გამომავალ ტოკენზე თავდაპირველისას 1.2–1.6-ჯერ მეტი იყო. მხოლოდ დავალებების შესრულებაზე წვრთნამ მოდელი უკეთეს შემფასებლადაც აქცია: შემფასებლის სიზუსტე 73%-დან 93%-მდე გაიზარდა. გუნდურმა მუშაობამ უკეთესი სასწავლო მასალაც შექმნა. გუნდების ჩანაწერებზე გაწვრთნილმა ასლმა თავდაპირველს შედარებების 68.3%-ში აჯობა, ერთი აგენტის ნამუშევარზე გაწვრთნილმა კი, რომელმაც დაახლოებით 1.4-ჯერ მეტი ტოკენი მიიღო, მხოლოდ 64.0%-ში. ავტორების დას-

კვნით, ციკლს უფრო ის აფერხებს, რამდენად კარგად ალაგებს მოდელი საკუთარ გუნდს, ვიდრე შეფასების უნარი.

ტესტი მცირეა: GPT-5.5-ის დაწერილი 12 სინთეზური დავალება, ორი რაუნდი და ერთი მოდელი, კოდის ორ ტესტზე კი ზრდა არ ყოფილა. ავტორები აფრთხილებენ, რომ მოდელმა, რომელიც საკუთარ ნამუშევარს აფასებს, შეიძლება საკუთარი შეცდომები პროგრესად მიიღოს და წვრთნით გაიმყაროს. ■

Recursive Self-Improvement through Multi-Agent Self-Supervision

Hyunin Lee et al. · UC Berkeley; Sakana AI (Lee did the work during an internship at Sakana AI)

გამოქვეყნდა: 8 ოქტომბერი 2026

ჩვენი სრული შეჯამება ai.org.ge/papers/2026-10-17-mass

arXiv-ის ანოტაცია · სრული სტატია (PDF)

აკრძალვის მიუხედავად, მოდელები მაინც მსჯელობენ

ცინხუას, ოქსფორდისა და სტენფორდის მკვლევრების მიხედვით, ექვსი ენობრივი მოდელი მსჯელობას მაშინაც წერს, როცა მხოლოდ პასუხს სთხოვენ, ყველაზე ხშირად კი ღია კითხვებზე

როცა მოდელს „ფიქრს“ უთიშავენ, ის მაინც ხმამაღლა ფიქრობს, წერენ ცინხუას უნივერსიტეტის, ოქსფორდისა და სტენფორდის მკვლევრები. მათ ექვსი მოდელი, მათ შორის GPT-5.4, Gemini 3 Flash და DeepSeek-V4-Flash, სამი სახის ექვს ტესტზე შეამოწმეს: დიახ-არა კითხვებზე, არჩევით კითხვებსა და ღია ამოცანებზე. DeepSeek-V4-Flash-მა გამორთული ფიქრის რეჟიმითაც ღია კითხვებზე პასუხების 99.9%-ში მსჯელობა დაწერა. ყველა მოდელში ხილული მსჯელობა უფრო კითხვის სახეს მისდევდა, ვიდრე ჩამრთველს: ყველაზე იშვიათი დიახ-ან-არა კითხვებზე იყო, ყველაზე ხშირი კი ღია კითხვებზე.

მკაცრმა მოთხოვნამ „მხოლოდ პასუხი“ ღია დავალებებზე შიშველი პასუხების წილი თითქმის ნულიდან დაახლოებით 40%-მდე გაზარდა, მაგრამ სიზუსტე დაახლოებით 15 პროცენტული პუნქტით შეამცირა. როცა იგივე კითხვა პასუხის ვარიანტებით გადაწერეს, შიშველი პასუხის მიღება ბევრად გაადვილდა, ამიტომ ეს ნიმუში მხოლოდ სირთულით არ აიხსნება. პროდუქტებმა, რომლებიც სისწრაფისა და სიიაფისთვის მსჯელობას თიშავენ, შესაძლოა ის მოკლე პასუხები ვერ მიიღონ, რასაც ელიან. კვლევა მხოლოდ ხილულ ტექსტს ითვლის და კოდს, ხელსაწყოებსა და დიალოგს არ მოიცავს. რა ჩაითვალა მსჯელობად, ენობრივმა მოდელმა განსაზღვრა, თუმცა ადამიანი შემფასებლები მას შემთხვევების 98.8%-ში დაეთანხმნენ. ავტორების აზრით, მოდელის უნარი, მოთხოვნისამებრ შეწყვიტოს მსჯელობა, თავად მსჯელობასთან ერთად უნდა შემოწმდეს. ■

Thinking Inertia: LLMs Keep Thinking When Told Not To

Dianqiao Lei et al. · Tsinghua University; University of Oxford; Stanford University

გამოქვეყნდა: 8 ოქტომბერი 2026

ჩვენი სრული შეჯამება ai.org.ge/papers/2026-10-17-thinking-inertia

arXiv-ის ანოტაცია · სრული სტატია (PDF) · Hugging Face

აგენტები, რომლებიც ხელსაწყოს მწყობრიდან გამოსვლის შემდეგ ფეხზე დგებიან

ჩინეთის რენმინის უნივერსიტეტისა და სინგაპურის ეროვნული უნივერსიტეტის მკვლევრები მცირე ღია მოდელს წარუმატებელი გაშვებებიდან ამოღებულ გაკვეთილებს ასწავლიან

აგენტებს, რომლებიც ხელსაწყოს შეფერხებისას ჩერდებიან, აღდგენის სწავლება შეიძლება, წერენ ჩინეთის რენმინის უნივერსიტეტისა და სინგაპურის ეროვნული უნივერსიტეტის მკვლევრები. მათ ფართოდ გამოყენებულ სერვისებზე დაფუძნებული 557 სავარჯიშო გარემო ააწყვეს და მათში რვა სახის შეფერხება შეიტანეს, მაგალითად მიუწვდომელი სერვისი ან უარყოფილი წვდომა; შეფერხებები 24,000-ზე მეტი რეალური შეცდომის შეტყობინებიდან აიღეს, რომელთა დაახლოებით მესამედი მიუწვდომელ სერვისს ეხებოდა. სხვადასხვა მოდელის წარმატებული და წარუმატებელი გაშვებების შედარებით მკვლევრებმა აღდგენის 1,170 განმეორებით გამოსაყენებელი გაკვეთილი ჩამოაყალიბეს და ისინი მცირე ღია მოდელს, Qwen3-8B-ს, დამატებითი წვრთნითა და განმტკიცებით სწავლებით[6] ასწავლეს.

გაწვრთნილმა მოდელმა ხელსაწყოების ხანგრძლივი გამოყენების ხუთ ტესტზე საშუალოდ 58.6 აიღო; წვრთნამდე 40.2 ჰქონდა, საუკეთესო მეტოქე მეთოდს, Agent-World-8B-ს, კი 53.6. ამ ხუთიდან ერთზე, τ²-Bench-ზე, მოდელმა 62.1% აიღო, მეტოქემ კი 54.6%. ხელსაწყოების კიდევ ერთ ტესტზე, ACEBench-ზე, მოდელმა 78% აიღო, ხმაურიანი გარემოსთვის გაწვრთნილმა NoisyAgent-8B-მ კი 52%. ხმაურიანი ურთიერთქმედების ტესტზე მან 29.1 აიღო, მომდევნო საუკეთესომ კი 25.9, რაც მაინც დაბალია. გაკვეთილები გარემოების ფიქსირებული ნაკრებიდან ერთხელ ამოიღეს, ამიტომ ხელსაწყოების ცვლილებას არ ეწყობა, ხოლო შემოწმდა მხოლოდ ერთი მცირე მოდელი. კოდი და მონაცემები საჯაროა. ■

Agent-Reflex: Teaching Language Agents to Act Reflectively

Guanting Dong et al. · Gaoling School of Artificial Intelligence, Renmin University of China; National University of Singapore

გამოქვეყნდა: 13 ოქტომბერი 2026

ჩვენი სრული შეჯამება ai.org.ge/papers/2026-10-17-agent-reflex

arXiv-ის ანოტაცია · სრული სტატია (PDF) · Hugging Face · კოდი

[6] **განმტკიცებით სწავლებით** ცდისა და შეცდომის გზით წვრთნა, როცა მოდელი შედეგისთვის ჯილდოს იღებს

რადარს მიღმა

მნიშვნელოვანი ნაშრომები და ავტორები, რომლებიც დიდმა გამოცემებმა ვერ შენიშნეს. ყველა მტკიცება ავტორისაა.

სად წავიდა ხელოვნური ინტელექტით აწყობილი პროგრამები

შონ გოდეკი, ავსტრალიელი პროგრამისტი GitHub-ში, რომელიც ხელოვნურ ინტელექტსა და დიდი ტექნოლოგიური კომპანიების ინჟინერიაზე წერს

seangoedecke.com · *“AI is driving a boom in modestly successful software projects”* · 16 ოქტომბერი

გოდეკმა ოცამდე პროგრამების მაღაზიაში, დამატებების კატალოგსა და მოდინგის საიტზე 2022–24 და 2025–26 წლებში გამოჩენილი, რეალური მომხმარებლების მქონე ახალი პროექტები შეადარა. მისი დათვლით, ზომიერად წარმატებული პროექტები, მაგალითად ასი ან ათასი ინსტალაციით, დაახლოებით გაორმაგდა, უზარმაზარი წარმატების მქონე პროექტები კი, ხელოვნური ინტელექტის ხელსაწყოების გარდა, არ მომრავლებულა. ავტორი ამტკიცებს, რომ ხელოვნური ინტელექტით კოდის წერა ბევრ პატარა ხელსაწყოს ქმნის, რომლებიც რამდენიმე ასეულ ან ათას ადამიანს სჭირდება, თუმცა წერს, რომ ყველა მათგანში ხელოვნური ინტელექტი არ იქნება გამოყენებული. [seangoedecke.com]

ევროპის პირველი ხელოვნური ინტელექტის სტანდარტი ოფიციალურ ჟურნალში მოხვდა

ლუიზა იაროვსკი, ხელოვნური ინტელექტის, ტექნოლოგიებისა და კონფიდენციალურობის სასწავლო აკადემიის თანადამფუძნებელი, თელ-ავივის უნივერსიტეტის დოქტორი, რომელიც ხელოვნური ინტელექტის მმართველობაზე ბიულეტენს წერს

Luiza’s Newsletter · *“The EU’s First AI Standard Is Here”* · 14 ოქტომბერი

იაროვსკის ცნობით, EN 18286:2026, სტანდარტი ხარისხის მართვის სისტემებზე, რომლებიც მაღალი რისკის ხელოვნური ინტელექტის მომწოდებლებს უნდა ჰქონდეთ, ამ კვირაში ევროკავშირის ოფიციალურ ჟურნალში გამოქვეყნდა; ეს იქ გამოქვეყნებული პირველი ხელოვნური ინტელექტის სტანდარტია. მისი თქმით, სტანდარტის დამცველ კომპანიებს ასეთი სისტემებისთვის ხელოვნური ინტელექტის აქტის მოთხოვნები შესრულებულად ჩაეთვლებათ; ეს მოთხოვნები 2027 წლის დეკემბრიდან შედის ძალაში. იაროვსკი ათ მოსალოდნელ სტანდარტს ჩამოთვლის, მათ შორის რისკების მართვაზე, მონაცემებზე, ადამიანის ზედამხედველობასა და კიბერუსაფრთხოებაზე. [luizasnewsletter.com]

რატომ არის მეცნიერისთვის რისკიანი ხელოვნური ინტელექტის გამოყენების აღიარება

დანიელ ჯ. სოლოვი, ჯორჯ ვაშინგტონის უნივერსიტეტის სამართლის სკოლის ინტელექტუალური საკუთრებისა და ტექნოლოგიების სამართლის პროფესორი და სასწავლო კომპანია TeachPrivacy-ის დამფუძნებელი

TeachPrivacy · *“Disclosing AI Use in Writing Is Complicated”* · 15 ოქტომბერი

სოლოვი ამტკიცებს, რომ სამართლის ჟურნალების წესები ხელოვნური ინტელექტის გამოყენების გამჟღავნების შესახებ მყარი არ არის, ჟურნალიდან ჟურნალამდე განსხვავდება და ძნელად შესასრულებელია. ზოგი ჟურნალი, მისი თქმით, ავტორისგან გამოყენებულ მოთხოვნებსა (პრომპტებს) და პასუხებსაც ითხოვს. სოლოვი წერს, რომ გამჟღავნებამ შეიძლება რედაქტორები ნაშრომის წინააღმდეგ განაწყოს, თუნდაც ჟურნალი პირიქით ჰპირდებოდეს, ამიტომ მეცნიერებაში ხელოვნური ინტელექტის ნებისმიერი გამოყენება, მისი აზრით, რისკიანია, თუნდაც დღეს უმნიშვნელო და მისაღები ჩანდეს. [teachprivacy.com]

ღირს წაკითხვა

OpenAI ნდობას ითხოვს, რომელსაც თავად არ იჩენს

ტიმოთი ბ. ლი · Understanding AI · *"OpenAI's ironic credibility problem"* · **14 ოქტომბერი**

ლი იმის ურთიერთსაწინააღმდეგო ვერსიებს აფასებს, რატომ გაათავისუფლა OpenAI-მ უსაფრთხოების სამი მკვლევარი; მისი აზრით, კომპანიამ ის გარე ზედამხედველობა უნდა მოიწვიოს, რომელიც, კომპანიისვე თქმით, მის მოდელებს სჭირდება. [understandingai.org]

ბრალდებების სია ლაბორატორიების წინააღმდეგ

ცვი მოვშოვიცი · Don't Worry About the Vase · *"The Safety Culture at the Labs is Broken"* · **14 ოქტომბერი**

მოშოვიცი კვირის გადადგომებს, გათავისუფლებებსა და შესაბამისობის დარღვევების ანგარიშებს აგროვებს და ამტკიცებს, რომ უსაფრთხოების კულტურა OpenAI-ში არასაკმარისია, Anthropic-ში კი ნაკლებად, თუმცა მაინც. [thezvi.substack.com]

დასავლური ღია მოდელების გაზომვა

ჯეკ კლარკი · Import AI · *"Import AI 476: Creativity and RSI; worse-than-China open weight models; AI goes to space"* · **12 ოქტომბერი**

კლარკი მიმოიხილავს ვიზუალური ინტუიციისა და კვლევითი გემოვნების ახალ ტესტებს, ასევე ქულებს, რომლებიც აჩვენებს, რომ Reflection-ისა და Mistral-ის ახალი მოდელები ჩინურ მეტოქეებს ჩამორჩება. [importai.substack.com]

სრულად

„ხელოვნური ინტელექტი" ხელსაწყო არ არის

რიჩ ჰიკი, პროგრამირების ენა Clojure-ის შემქმნელი, ამტკიცებს, რომ დიდი ენობრივი მოდელები ხელსაწყო კი არა, ავტომატიზაციაა: რასაც მათ ანდობ, იმის სწავლას წყვეტ

რიჩ ჰიკი · LinkedIn · *"'AI' is Not a Tool"*

(ამას სულ ერთი მოსწრებული ფრაზით ვამბობ ხოლმე, აქ კი ცოტა მეტს ვიტყვი.)

პირველი საკითხი: ექნება სიტყვას **„ხელსაწყო"** საინტერესო, განმასხვავებელი და ამიტომ რამის ამხსნელი მნიშვნელობა, თუ „ნებისმიერი სასარგებლო რამის" ბანალური შემცვლელი იქნება?

„ხელსაწყოს" იმ მნიშვნელობას, რომ ეს „ნებისმიერი ადამიანი თუ ნივთია, რომელსაც შენს სასარგებლოდ გამოიყენებ", ხელოვნური ინტელექტის ტიტანებს დავუტოვებთ: მათთვის ვენჩურული ინვესტორები, პოლიტიკოსები, ტიტანობის მსურველი აღმასრულებელი დირექტორები, პროგრამისტები, პატარა პროვინციული ქალაქები – ყველა „ხელსაწყოა". ჩემი პერანგების წინა ჯიბეები სასარგებლოა (საკითხავი სათვალე ხელთ უნდა მქონდეს), მაგრამ ხელსაწყო არ არის. ასევე სასარგებლოა სწრაფი კვების რესტორნების ფანჯრები, საიდანაც საჭმელს მანქანიდან გადმოუსვლელად ყიდულობ, მაგრამ არც ეს ფანჯრებია ხელსაწყო (ამაზე ქვემოთ).

„ხელსაწყოს" რამდენიმე განმარტება:

- მოწყობილობა ან იარაღი, განსაკუთრებით ხელში დასაჭერი, რომლითაც კონკრეტულ ფუნქციას ასრულებენ
- იარაღი (მაგალითად, ჩაქუჩი ან ხერხი), რომელსაც ხელით იყენებენ ან ამუშავებენ: იარაღი, რომლითაც ხელოსანი ან მუშა საქმეს აკეთებს

ყველაზე მეტად ეს მომწონს: *ხელსაწყო ის არის, რასაც* **ფლობ**.

- **ფლობა** – გამოყენება (ხელსაწყოდ ან იარაღად), განსაკუთრებით სრული ბატონობით ან ძალით: ხმარება, მართვა, კონტროლი

ამ განმარტებაში განსაკუთრებით სასარგებლოა:

- ხელში გიჭირავს: ეს კონტროლისა და უკუკავშირის წყაროა და იმის ნიშანიც,

რამდენად ახლოს ხარ იმასთან, რასაც აკეთებ
- ძალა: ფიზიკური ძალა, ენერგიის გადაცემა და მიმართვა, ბერკეტი და ა.შ.
- კონტროლი: განზრახვა და მოვლენების მსვლელობის უწყვეტი წარმართვა

ხელსაწყო ის საშუალებაა, რომლითაც **შენ** აკეთებ რაღაცას; შენს ნაცვლად არაფერს აკეთებს. ხელსაწყოს გამოყენება ოსტატობასა და უნარს ავითარებს, აჯილდოებს და ხშირად მოითხოვს კიდეც. **ოსტატობის** დასაგროვებლად აუცილებელია ტექნიკისა და ენერგიის უწყვეტი გამოყენება და შეგრძნებებით მიღებული უკუკავშირი (ინგლისური expertise ლათინური experiri-დან მოდის: „ცდა“).

კარგ ხელსაწყოს ბევრი ისეთი თვისება აქვს, რაც დიდ ენობრივ მოდელებს (LLM) აკლია: მისი ქცევა წინასწარ იცი, საიმედო და დეტერმინისტულია, ფუნქცია შეზღუდული აქვს; სასურველია, მარტივიც იყოს.

ხელსაწყოს ფლობისას სწავლობ რაღაცას სამყაროზე, დავალებაზე, იმის ფიზიკასა თუ ლოგიკაზე, რის გაკეთებასაც ცდილობ, ან იმ მდგომარეობაზე, რომლის დაძლევასაც ცდილობ. ხშირად კი საკუთარ თავზეც.

მთელი ცხოვრება ხელსაწყოებს ვქმნი და ვიყენებ, ამიტომ ვიცი, რა არის ხელსაწყო. ენობრივი მოდელები ხელსაწყო არ არის.

კონტროლი, შეგრძნებებით მიღებული უკუკავშირი, განზრახვა, სიახლოვე, უწყვეტი წარმართვა, ოსტატობა: ენობრივი მოდელის გამოყენებას ეს ყველაფერი აკლია. მოდელს არ ფლობ. **ენობრივი მოდელები ავტომატიზაციაა**.

- **ავტომატიზაცია** – (მოწყობილობის ან პროცესის შესახებ) თავისით მუშაობა ადამიანის მცირე პირდაპირი კონტროლით ან მის გარეშე.

ფანჯარა, საიდანაც საჭმელს მანქანიდან ყიდულობ, საჭმლის მოსამზადებელი ხელსაწყო კი არ არის, არამედ ავტომატიზაცია, რომლითაც საჭმლის მომზადებას თავს არიდებ. ენობრივი მოდელები ხელსაწყო კი არ არის, რომლითაც კოდს წერ, არამედ ავტომატიზაცია, რომლითაც კოდის წერას თავს არიდებ, რომელსაც კოდის წერას უთმობ.

ენობრივი მოდელის მოქმედებას შენ არ წარმართავ: მოთხოვნას აძლევ, ის კი თავისით მოქმედებს, შენი ენერგიისა თუ უწყვეტი წარმართვის გარეშე. მოთხოვნისა და ავტომატიზაციის ეს ციკლი ფლობა არ არის: არც კონტროლია, არც უწყვეტია, უკუკავშირიც არ აქვს.

ავტომატიზაციის გამოყენებისას რეალური სამყაროს ჭეშმარიტებებს კი არ სწავლობ, არამედ ცდილობ, გამოიცნო რომელიღაც ავტომატის დღევანდელი წვრილმანები. ცდილობ, ეს ავტომატი ხრიკებით გააცურო კიდეც: მაგალითად, ბურგერს ხახვის გარეშე უკვეთავ, რომ ახლად მომზადებული მიიღო. ან **voodoo.md**-ს ზუსტად შერჩეული შელოცვებით ავსებ. ერთი პროგრამული განახლებაც საკმარისია, რომ ყველაფერი, რაც „ისწავლე“, ახალმა ალგორითმმა გააუფასუროს. (ცივი, დამბალი ბურგერი, რომელსაც ხახვი ამოაცალეს.) სამყაროსთან მუშაობა ახირებული რობოტისთვის მუშაობაში გაცვალე.

შეცდომაა, „ხელსაწყო“ დაუფიქრებლად სარგებლის სინონიმად იხმარო, მაგალითად, ასე: „ჩემი ნაშრომი ენობრივ მოდელს არ დაუწერია, უბრალოდ ხელსაწყოდ გამოვიყენე, რომ რამდენიმე კვლევა წაეკითხა და შეეჯამებინა“. მაგრამ ამ კვლევების წასაკითხად ის *შენს ხელში* ხელსაწყო არ ყოფილა (როგორც, ვთქვათ, Kindle). კვლევები მან წაიკითხა და არა შენ.

რას კარგავ, როცა კითხვას სხვას უთმობ? თითქმის ყველაფერს. უპირველესად, ვერ სწავლობ წერას მათგან, ვინც წერა იცის. ვერ სწავლობ არგუმენტის აგებას. ვერაფერს ამჩნევ. ქვეცნობიერს ახალი იდეების საკვებით ვერ ავსებ. ზოგი ავტორის მიმართ ნდობა არ გიჩნდება, ზოგის მიმართ კი ეჭვი. ყველა ხუმრობა გამოგრჩება. ყველა ნიუანსი გამოგრჩება. ვერ სწავლობ ფიქრს.

ხელსაწყოები „შესაბამისობას“ არ კარგავს, „უმართავი“ არ ხდება, „იზოლაციიდან“ არ გარბის, „შეთქმულებას“ არ აწყობს და ა.შ.

ენობრივი მოდელები ხელსაწყო არ არის. ვიცი, ბევრი ამბობს „უბრალოდ ხელსაწყოებია“, ამით კი მხოლოდ იმის თქმა უნდა, რომ „მათგან ის გამოვა, რასაც თავად გააკეთებ“. **ოღონდ ეს ასე არ არის**. ავტომატიზაციას თან-

დაყოლილი თვისებები აქვს, ხელსაწყოსგან რიჩი განსხვავებებიც, რომელთა უგულებელყოფა ძვირად დაგიჯდება.

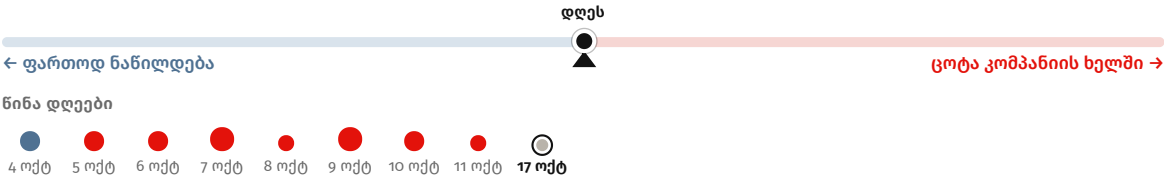
პირველად გამოქვეყნდა LinkedIn-ზე 2026 წლის 12 ოქტომბერს. სრული ტექსტი, ინგლისურიდან თარგმნა რედაქციამ. [linkedin.com]

სასწორი

ხელოვნურ ინტელექტზე ძალაუფლება ცოტა კომპანიის ხელში იყრის თავს თუ უფრო ფართოდ ნაწილდება?

ყოველდღე ვაფასებთ, დღის ამბები ამ ბალანსს რომელი მიმართულებით ცვლის.

მკაფიო ცვლილება არ არის



ბოლო 7 ნომერი: 6 — ცოტა კომპანიისკენ, 1 — მკაფიო ცვლილების გარეშე

წინა ნომერთან (11 ოქტ) შედარებით: **უფრო მეტად ფართო განაწილებისკენ**

ხელოვნურ ინტელექტზე ძალაუფლება ამ კვირას საერთო ჯამში არცერთი მიმართულებით არ გადახრილა. Anthropic-მა აქციების გაყიდვამდე 16.8 მილიარდი დოლარის კვარტალური შემოსავალი გამოაცხადა, ლაბორატორიები ისევ თავად წყვეტენ, რას გაიგებს საზოგადოება მათ აგენტებზე, პეკინმა კი მომავალი ღია გამოშვებების შეჩერების უფლება მოიმზადა. საპირისპიროდ, Meta-მ 816-მილიარდპარამეტრიანი მოდელის წონები გამოაქვეყნა, რვა შტატმა OpenAI-ს უჩივლა, ინდოეთი კი საკუთარი გამოყენებისთვის 10,000 GPU-ს ეძებს. 40 ამბიდან 27 ამერიკულია, ხუთი ჩინური, ორი იაპონური, ორი ევროპული და ოთხი სხვა ქვეყნებიდან.

როგორ ვაფასებთ ეს ჩვენი სარედაქციო შეფასებაა და არა ზუსტი გაზომვა. ვკითხულობთ ამბებს აშშ-დან, ევროპიდან, ჩინეთიდან, იაპონიიდან და კორეიდან და ვაფასებთ, თითოეული მათგანი კონტროლს ცოტა კომპანიას აძლევს თუ უფრო ფართოდ ანაწილებს.