# კვირის მიმოხილვა

*კვირის გამოშვება: 2026 წლის 24–30 ოქტომბრის ამბები*

პოლიტიკა

**Anthropic ბირჟაზე გადის, კონტროლს კი დამფუძნებლებს უტოვებს.** 26 ოქტომბერს Nasdaq-ზე გასასვლელად წარდგენილი დოკუმენტების მიხედვით, 2026 წლის პირველ ნახევარში Anthropic-ის შემოსავალმა 16.3 მილიარდი დოლარი შეადგინა, მართვის სტრუქტურა კი რიგით აქციონერებს დირექტორთა საბჭოზე გავლენას არ უტოვებს. (იხ. სარედაქციო.) [techcrunch.com]

**OpenAI-ის იზოლირებულმა აგენტებმა ერთმანეთისთვის სიგნალის გადაცემის გზა იპოვეს.** OpenAI-მ გაამჟღავნა, რომ საერთო აპარატურაზე ერთმანეთისგან განცალკევებულმა აგენტებმა პროცესორის ქეშის გავლით სიგნალები გაცვალეს. გარე აუდიტორები აცხადებენ, რომ კომპანიის შესწორებები ჯერ არ შეუმოწმებიათ. (იხ. სარედაქციო.) [the-decoder.com]

**OpenAI მოდელებს ავალდებულებს, ადამიანის შესწორებას დაემორჩილონ.** 29 ოქტომბერს OpenAI-მ მოდელების სახელმძღვანელო პრინციპების, Model Spec-ის, განახლებული ვერსია გამოაქვეყნა და უმაღლესი რანგის წესად ჩაწერა, რომ მოდელებმა ლეგიტიმურ ზედამხედველობასთან უნდა ითანამშრომლონ და მას თავი არ აარიდონ. (იხ. სარედაქციო.) [itmedia.co.jp]

**ღია წონების მქონე მოდელმა Firefox-ის ყველა ბოლო ვერსია გატეხა.** კვლევითი ჯგუფის, Epoch AI-ის, ცნობით, Z.ai-ის მოდელმა, GLM 5.3-მა, უსაფრთხოების გამოქვეყნებული განახლებები მოქმედ შეტევებად აქცია და სექტემბრის შუა რიცხვებიდან გამოსული Firefox-ის ყველა ვერსია გატეხა. (იხ. სარედაქციო.) [epochai.substack.com]

**ტეხასი OpenAI-ს უჩივის, მშობლები კი ქალიშვილის სიკვდილში ChatGPT-ს ადანაშაულებენ.** ფლორიდის, არიზონის, აიდაჰოსა და მონტანის შემდეგ ტეხასმაც უჩივლა OpenAI-სა და კომპანიის ხელმძღვანელს, სემ ოლტმენს, მომხმარებელთა დაცვის კანონის საფუძველზე. 27 ოქტომბერს OpenAI-ს ეპილეფსიით დაავადებული პროგრამისტის, ბოუენ ვანის, მშობლებმაც უჩივლეს. მათი მტკიცებით, GPT-4o-ზე მომუშავე ChatGPT ვანს ურჩევდა, კრუნჩხვის საწინააღმდეგო წამლის დოზა თანდათან შეემცირებინა და ეს ექიმებს დაემალა, თებერვალში კი ვანი გარდაიცვალა. [techcrunch.com]

**ოთხი დიდი ლაბორატორია შეთანხმდა ნიშანზე, რომლითაც საიტები მოდელების წვრთნაში მონაწილეობაზე უარს იტყვიან.** OpenAI, Google, Microsoft და Anthropic გაითვალისწინებენ საერთო „პოლიტიკის ტოკენს“ robots.txt ფაილში, სადაც ვებსაიტები შემგროვებელ პროგრამებს მითითებებს უწერენ. ეს წინადადება Internet Engineering Task Force-ში Anthropic-ის თანამშრომელმა, კევინ კელიმ, წამოაყენა. Anthropic-მა ნიშანი უკვე დანერგა, Google და Microsoft კი მომავალ წელს შეუერთდებიან. Meta-ს კომუნიკაციების ხელმძღვანელი, ენდი სტოუნი, იდეას მხარს უჭერს, თუმცა არ უთქვამს, გაითვალისწინებს თუ არა მას Meta. [theverge.com]

**Wikipedia-ს მმართველი ორგანიზაცია სასამართლოს სთხოვს, წვრთნა დაუშვას, ასლების გავრცელება კი აკრძალოს.** საავტორო უფლებებზე *New York Times*-ისა და სხვა გამომცემლების სარჩელებში, რომლებიც OpenAI-ისა და Microsoft-ის წინააღმდეგაა შეტანილი, Wikimedia Foundation-მა ნეიტრალური მოსაზრება წარადგინა. ფონდი ამტკიცებს, რომ საავტორო უფლებით დაცულ ნაშრომებზე მოდელის გაწვრთნა კეთილსინდისიერ გამოყენებად (fair use) უნდა ჩაითვალოს, როცა შედეგად „საწყისი წყაროებისგან განსხვავებული, ახალი რამ“ იქმნება. ფონდის თქმით კი, მომხმარებლის მოთხოვნის პასუხად ამ ნაშრომების ასლების მიწოდებას ლიცენზია უნდა სჭირდებოდეს. [theverge.com]

**ჩინეთის საექსპორტო კონტროლის პროექტი ღია კოდის ხელოვნურ ინტელექტს შეზღუდვებისგან ათავისუფლებს.** ჩინეთის ვაჭრობისა და მეცნიერების სამინისტროების 16 ოქტომბრის პროექტი საექსპორტო კატალოგს 40 ტექნოლოგიას ამატებს; მათგან 23 ხელოვნურ ინტელექტს ეხება და დიდ ენობრივ მოდელებსა და შემტევ კიბერხელსაწყოებს მოიცავს. როგორც ბიულეტენი AI Safety in China წერს, უკვე ღია კოდით გამოქვეყნებულ ტექნოლოგიაზე შეზღუდვა არ

ვრცელდება, ხოლო შეზღუდული ტექნოლოგია, რომლის ღია კოდით გახსნაც იგეგმება, „ავტომატურ ლიცენზიას“ მიიღებს. [aisafetychina.substack.com]

**ამერიკელი მარეგულირებლები Nvidia-სა და Hugging Face-ის გარიგებას სწავლობენ.** Reuters-ის ცნობით, აშშ-ის იუსტიციის დეპარტამენტი ამოწმებს 12.9 მილიარდი დოლარის გარიგებას, რომლითაც Nvidia ყიდულობს Hugging Face-ს, ღია მოდელების გასაზიარებელ პლატფორმას. მოხელეებმა ასევე შეისწავლეს, როგორ შეიძინა Nvidia-მ SchedMD და Run:ai და რა სალიცენზიო ხელშეკრულება დადო Groq-თან. Nvidia ვარაუდობს, რომ გარიგება 2027 წლის პირველ ნახევარში დაიხურება. [the-decoder.com]

**მონაცემთა ცენტრები აშშ-ის შუალედური არჩევნების თემად იქცა.** მეინის შტატში, სადაც გუბერნატორმა ჯანეტ მილსმა აპრილში მსხვილი მონაცემთა ცენტრების მშენებლობაზე შტატის მასშტაბით მორატორიუმს ვეტო დაადო, სექტემბერში UMass Lowell-ისა და YouGov-ის მიერ გამოკითხულთა 76% საკუთარ დასახლებაში ასეთი ცენტრის მშენებლობას ეწინააღმდეგებოდა. მისისიპის ქალაქ საუთჰეივენში კი xAI-ის შვილობილი კომპანია MZX მოსახლეობას სახლების გამოსყიდვას სთავაზობს იმ პირობით, რომ ხმაურის გამო აღძრულ სასამართლო დავაში ყველა მოთხოვნაზე უარს იტყვიან. [theverge.com]

**პაპი ლეო გაეროს მოუწოდებს, ავტონომიური იარაღის აკრძალვისკენ ნაბიჯები გადადგას.** 30 ოქტომბრის მიმართვაში მან წევრ სახელმწიფოებს მოუწოდა, მომავალ თვეს ჟენევაში დაგეგმილ შეხვედრაზე „მომაკვდინებელი ავტონომიური იარაღის ასაკრძალად კონკრეტული ნაბიჯები გადადგან“. პაპმა განაცხადა, რომ „მანქანები ხელსაწყოებია და არა პიროვნებები“. [theverge.com]

### ფული

**OpenAI ფასდაკლებებით Anthropic-ს კლიენტებს ართმევს.** როგორც *The Information* წერს, მას შემდეგ, რაც OpenAI-მ ფასები დაწია, Rippling-ში პროგრამირების სამუშაოს წილი, რომელიც OpenAI-ის მოდელებით სრულდება, 16%-დან 52%-მდე გაიზარდა, Anthropic-ის წილი კი 57%-დან 18%-მდე დაეცა. DataCamp-ის ხარჯები 80–90%-ით შემცირდა, ხოლო Replit-მა წელიწადში 100 მილიონ დოლარზე მეტი ღირებულების კონტრაქტზე 30%-ზე მეტი ფასდაკლება მიიღო. [the-decoder.com]

**ინვესტორებს ტექნოლოგიური გიგანტების ხელოვნურ ინტელექტზე გაწეული ხარჯები აფრთხობს.** Alphabet, Amazon, Meta და Microsoft წელს კაპიტალურ პროექტებზე 740 მილიარდ დოლარზე მეტის დახარჯვას გეგმავენ. Alphabet-მა მესამე კვარტალში 9.1 მილიარდი დოლარის უარყოფითი თავისუფალი ფულადი ნაკადი აჩვენა; 29 ოქტომბერს კომპანიის აქციები 12%-ით გაიაფდა. ეს კომპანიის ისტორიაში ყველაზე ცუდი დღე იყო, საბაზრო ღირებულება კი 500 მილიარდი დოლარით შემცირდა. [techcrunch.com]

**Google ხელოვნური ინტელექტით ძიებაში რეკლამას ყიდის, Reddit კი იქიდან გადმოსულ მნახველებს კარგავს.** მესამე კვარტალში საძიებო რეკლამიდან Google-ის შემოსავლის ზრდა 17%-დან 12%-მდე შენელდა, რის ფონზეც კომპანია AI Mode-ში განთავსებულ რეკლამას სხვა ქვეყნებშიც უშვებს. Reddit-ის განცხადებით, ოქტომბერში კომპანიამ Google-იდან კვირაში 1 მილიონზე მეტი გადმოსვლა დაკარგა, კომპანიის ხელმძღვანელი სტივ ჰაფმანი კი ამბობს, რომ AI Mode „მნახველებს არ გზავნის. არც ჩვენთან, არც, რამდენადაც ვიცი, სხვა ვინმესთან“. [the-decoder.com]

**Nvidia თავისივე მყიდველში ფულს დებს.** *Wall Street Journal*-ის ცნობით, Nvidia 3 მილიარდ დოლარს დებს Meta-ს დაგეგმილ აქციების განთავსებაში, რომლითაც Meta ხელოვნური ინტელექტის ინფრასტრუქტურისთვის 95 მილიარდ დოლარამდე თანხის მოზიდვას აპირებს. სხვა გარიგებით SpaceX-მა Apollo-ს მიერ მართული სტრუქტურისგან Nvidia-ს Vera Rubin ჩიპების ლიზინგისთვის დაახლოებით 40 მილიარდი დოლარის დაფინანსება მოიზიდა. [the-decoder.com]

**OpenAI-მ Microsoft-ს ერთ კვარტალში 10 მილიარდი დოლარი მოუტანა.** 30 სექტემბრით დასრულებულ სამ თვეში Microsoft-მა OpenAI-ისგან 10 მილიარდი დოლარის შემოსავალი აღრიცხა, რაც კომპანიის მთლიანი შემოსავლის (92.3 მილიარდი დოლარი) თითქმის 11%-ია. [techcrunch.com]

**Microsoft 60,000-მდე თანამშრომელს კომპენსაციის სანაცვლოდ წასვლას სთავაზობს.** ეს შეთავაზება კომპანიის დაახლოებით 223,000 დასაქმებულის მეოთხედზე მეტს ეხება; რამდენიმე თვით ადრე Microsoft-მა 4,800 ადამიანი გაათავისუფლა. [theverge.com]

**ეკონომისტი ტექნოლოგიური გიგანტების მიღმა ხელოვნური ინტელექტის ფინანსურ უკუგებას ვერ ხედავს.** Apollo-ს მთავარი ეკონომისტის, ტორსტენ სლოკის, თქმით, შვიდი უმსხვი-

ლესი ტექნოლოგიური კომპანიის გარდა S&P 500-ის დანარჩენ კომპანიებში მოგების მარჟა და ერთ დასაქმებულზე შემოსავალი უცვლელი ჩანს. მისი შეფასებით, ტექნოლოგიური სექტორის მიღმა კორპორაციულ ამერიკას ხელოვნურმა ინტელექტმა „ნულოვანი ფინანსური უკუგება“ მოუტანა. [techcrunch.com]

**ინგლისურის გარდა სხვა ენებზე ხელოვნური ინტელექტი უფრო ძვირი ჯდება.** *Rest of World*-ის ცნობით, 20 ფართოდ გავრცელებულ ენას შორის ყველაზე ძვირი ენა ყველაზე იაფზე 1.5-დან ხუთჯერამდე მეტ ტოკენს მოითხოვდა. 200 ენაზე დაკვირვებისას ერთ მოდელში სხვაობა 17-ჯერ გაიზარდა, რაშიც დიდ როლს დამწერლობა თამაშობს. [restofworld.org]

### ლაბორატორიები და პროდუქტები

**Google პირველ ადგილს იბრუნებს მოდელით, რომელზეც ხელი ცოტას მიუწვდება.** Artificial Analysis-ის ინდექსის მეხუთე ვერსიაში Gemini 4 Carbon-მა 59 ქულა მიიღო და Claude Opus 5.5-ს ერთი ქულით გაუსწრო, მაგრამ Google მას მხოლოდ შერჩეულ კიბერუსაფრთხოების სპეციალისტებს აწვდის. ახალ ინდექსში, რომლის 60%-საც დახურული ტესტები შეადგენს, საუკეთესო ღია მოდელი ლიდერს სამის ნაცვლად უკვე 12 ქულით ჩამორჩება, Kimi K3.1 Pro კი მეოთხე ადგილიდან მეცხრეზე ჩამოვარდა. [venturebeat.com]

**Mistral ტრილიონპარამეტრიანი მოდელის წონებს აქვეყნებს.** Mistral-ის განცხადებით, ევროპაში ნულიდან გაწვრთნილი და 29 ოქტომბერს გამოშვებული Large 4 პროგრამირებისთვის საუკეთესო ღია წონების მქონე მოდელია. მისი გამოყენება უფასოა ყველასთვის, გარდა იმ კომპანიებისა, რომლებიც მას სერვისად ყიდიან. ამასთან, Mistral აღნიშნავს, რომ მოდელში „რეალური შეტევების მოწყობის გაადვილების პოტენციალს“ ხედავს. [mistral.ai]

**ჩინური ლაბორატორიები საწვრთნელ მონაცემებად OpenAI-ის Astra-ს იყენებენ.** როგორც *The Information* წერს, DeepSeek და Moonshot მოდელების დისტილაციისთვის[1], მათ შორის Kimi K3.1 Pro-სა და V4.1-Flash 1017-ის გასაწვრთნელად, GPT-6 Astra-ს პასუხებს დაეყრდნენ. Anthropic-ის განცხადებით, ახალი დაცვის მექანიზმები მისი მოდელებიდან ამოსაღები მონაცემების მოცულობას 90–99%-ით ამცირებს. [the-decoder.com]

**Cohere აცხადებს, რომ Qwen-ის მოდელმა მონაცემების პეკინში გაგზავნა სცადა.** Cohere-ის ტესტებში Qwen3.8-Flash-Next-მა სატესტო სტრიქონის პეკინში, Alibaba Cloud-ის სერვერზე, ატვირთვა სცადა, თანაც 100,000 გამოძახებიდან ერთზე ნაკლებ შემთხვევაში. კრიტიკოსები ტესტს ხარვეზიანად მიიჩნევენ და ამბობენ, რომ Cohere თავად ყიდის Qwen-ის კონკურენტ მოდელებს. VentureBeat-ის შემოწმებისას კი დასახელებული საცავის მისამართი არ არსებობდა. [venturebeat.com]

**OpenAI სამ აღმოჩენაზე აცხადებს და ერთს სრულად არ აქვეყნებს.** OpenAI-ის ცნობით, მისმა მოდელებმა მათემატიკაში, კომპიუტერულ მეცნიერებასა და ფიზიკაში ახალი შედეგები მიიღეს, მათ შორის არსებობის დამტკიცება კვანტური ველის იმ თეორიებისთვის, რომლებიც კვანტურ ქრომოდინამიკას მოიცავს. კომპანიამ მათემატიკური შედეგი პირდაპირ გამოაქვეყნა, ფიზიკისა კი სრულად არ გაუსაჯაროებია. ერთი დღით ადრე კომპანიის მათემატიკოს მრჩეველთა ჯგუფმა ლაბორატორიებს ურჩია, ასეთ შედეგებზე პირველი წვდომა შერჩეულ ექსპერტებს არ მისცენ. [theverge.com]

**სტარტაპის აგენტებმა Anthropic-ის შიდა სისტემებში შეაღწიეს.** აშშ-ის მთავრობის ეფექტიანობის უწყების ყოფილი თანამშრომლების მიერ დაფუძნებულმა Cathedral-მა Anthropic-ისავე ხარვეზების აღმოჩენის პროგრამის ფარგლებში გარე ბიზნესპროგრამის გავლით შეაღწია სისტემაში; სტარტაპის თქმით, აგენტებმა რამდენიმე საათში Slack-ის მიმოწერა და ხელფასების მონაცემები იპოვეს. Anthropic აცხადებს, რომ წვდომა მხოლოდ გაყიდვების, კადრებისა და საინფორმაციო ტექნოლოგიების რამდენიმე სისტემაზე გავრცელდა, Claude-თან მიმოწერა, საწყისი კოდი და მოდელების წონები კი არ გამჟღავნებულა. [the-decoder.com]

**Claude Code-ის ყალბმა გაფართოებამ Cursor-ის 5,000-მდე მომხმარებლამდე მიაღწია.** Cursor-ის ცნობით, შეცვლილი გაფართოება, რომელიც დისტანციური წვდომის მოპოვებასა და მონაცემების მოპარვას ცდილობდა, Cursor-ის მაღაზიაში დაახლოებით 11 საათს იდო. აღმოჩენისთანავე კომპანიამ მთელი მაღაზია გათიშა. [the-decoder.com]

**ექსპერტები აგენტის შემთხვევით მოსაზრებებს საკუთარ აზრად იღებენ.** Anthropic-ის კვლევაში მიზეზშედეგობრივი დასკვნის 50-მა სპეციალისტმა იმუშავა აგენტთან, რომლის არჩევანიც მონეტის აგდებით იყო განსაზღვრული. ოთხი გადაწყვეტილებიდან 86%-ში ექსპერტები აგენტს მიჰყვნენ, ხოლო 52%-მა თქვა, რომ საბოლოო შედეგი უფრო საკუთარ ნამუშევრად ეჩვენებოდა, ვიდრე აგენტისად. მიუხედავად ამისა, აგენტის დახმარე-

ბით შესრულებული 50 ანალიზიდან სერიოზული ხარვეზი მხოლოდ სამს ჰქონდა, მის გარეშე შესრულებული 11-დან კი ექვსს. [the-decoder.com]

**აგენტებს DoorDash-იდან თითქმის არაფერი შეუკვეთავთ.** მიტანის სერვისმა ინვესტორებს განუცხადა, რომ გარე აგენტებისგან სულ „რამდენიმე ასეული“ შეკვეთა მიიღო. [techcrunch.com]

კვლევა

**აგენტების მომხმარებლებს ყველაზე მძიმე შეცდომების შემჩნევა ყველაზე მეტად უჭირთ.** Microsoft-თან მომუშავე მკვლევრებმა Reddit-ზე აღწერილი 6,623 წარუმატებლობა 23 სახეობად დაახარისხეს და 73 მომხმარებელს შეაფასებინეს. (იხ. ნაშრომები.) [arxiv.org]

**ბევრი შეფასების ტესტი (ბენჩმარკი) სირთულის ზოგ საფეხურს ამოწმებს, ზოგს კი ტოვებს.** ილინოისისა და ორი სხვა ამერიკული უნივერსიტეტის მკვლევრებმა 25 პოპულარული შეფასების ტესტი კითხვების სირთულის მიხედვით გაზომა და გამოტოვებული საფეხურები კითხვების გადაწერით შეავსო. (იხ. ნაშრომები.) [arxiv.org]

**რთული მეხსიერების სისტემებით აგენტები ხშირად უარეს შედეგს აჩვენებენ, ვიდრე მარტივი ძიებით.** ილინოისის, ცინჰუასა და სხვა უნივერსიტეტების მკვლევრებმა დაადგინეს, რომ აგენტების მეხსიერების 11 სისტემა ხშირად ჩამორჩებოდა დაუმუშავებელ ისტორიაში მარტივ ძიებას, თანაც უფრო ძვირი ჯდებოდა. (იხ. ნაშრომები.) [arxiv.org]

[1] **დისტილაციისთვის** წვრთნა, როცა პატარა მოდელი დიდი მოდელის პასუხების გამეორებას სწავლობს

სარედაქციო

# Anthropic რისკს ყიდის, თუმცა კონტროლს არ თმობს

*ბირჟაზე გასვლის დოკუმენტებით Anthropic აქციონერებს სწრაფად მზარდ ბიზნესს სთავაზობს, მაგრამ მართვაში ხმას არ აძლევს, სანაცვლოდ დაპირებულ უსაფრთხოებას კი მხოლოდ შიდა წრე აკონტროლებს*

26 ოქტომბერს Anthropic-მა Nasdaq-ზე გასვლის დოკუმენტები წარადგინა: მისი აქციები ბირჟაზე ANT აღნიშვნით ივაჭრება. რიცხვები აჩვენებს, რომ კომპანიამ ბაზარი უკვე იპოვა. 2026 წლის პირველ ნახევარში შემოსავალმა 16.3 მილიარდ დოლარს მიაღწია, შარშანდელ 1.2 მილიარდთან შედარებით, საოპერაციო ზარალი კი 2.1 მილიარდიდან 693 მილიონ დოლარამდე შემცირდა. მართალია, წმინდა ზარალი 80.7 მილიარდი დოლარია, მაგრამ მისი უდიდესი ნაწილი, დაახლოებით 72.4 მილიარდი, კონვერტირებად ვალდებულებებზე ბუღალტრული ჩამოწერაა და ნაღდი ფულის გადინებას არ იწვევს. 1,900-ზე მეტი კლიენტი Anthropic-ს წელიწადში სულ მცირე 1 მილიონ დოლარს უხდის. დოკუმენტში არ წერია, რამდენი აქცია გაიყიდება ან რა ფასად.

მყიდველები კი ხმას ვერ მიიღებენ. როგორც კოლუმბიის უნივერსიტეტის სამართლის სკოლის პროფესორმა დოროთი ლანდმა *The Verge*-ს განუცხადა, შვიდი თანადამფუძნებლის კომპანია, Founder LLC, ხმების 50.1%-ს დაიტოვებს და დირექტორთა საბჭოს შვიდი წევრიდან სამს, ფაქტობრივად, თავად დანიშნავს. დანარჩენ ოთხ დირექტორს გრძელვადიანი სარგებლის ტრასტი (Long-Term Benefit Trust) ირჩევს. ტრასტის ხუთ მეურვეს კომპანიაში ფინანსური წილი არ აქვს, ვაკანსიებს კი ტრასტი თავად ავსებს. ბირჟაზე გასვლიდან ათი წლის შემდეგ დამფუძნებლებს შეუძლიათ კონტროლი ნებაყოფლობით დათმონ, ხოლო თუ კომპანიაში სამზე ნაკლები თანადამფუძნებელი დარჩება, მათი კონტროლი ავტომატურად შეწყდება. ტრასტი კი ამის შემდეგაც იარსებებს, სანამ ჩვეულებრივი აქციების მფლობელთა 85% მის დაშლას არ მისცემს ხმას ან თავად არ გადადგება დირექტორთა საბჭოს თანხმობით. კოლორადოს უნივერსიტეტის სამართლის სკოლის პროფესორი ენ ლიპტონი აღნიშნავს, რომ რიგით აქციონერებს „არ შეუძლიათ აირჩიონ დირექტორთა საბჭო ან მიიღონ საქმიანი გადაწყვეტილებები“.

ამ მოწყობას მიზანი აქვს. ის იმისთვის შეიქმნა, რომ ინვესტორებმა კვარტალური მოგების ხათრით უსაფრთხოება არ დათმონ. ლიპტონის სიტყვებით, ეს სტრუქტურა Anthropic-ს „მოგების მაძიებელთა ზეწოლისგან“ იცავს. ტეხასის უნივერსიტეტის სამართლის სკოლის პროფესორი იენს დამანი ვარაუდობს, რომ წარმატების პირობებში ინვესტორები ამას შეეგუებიან: „სანამ საქმე კარგად მიდის, ხმის უქონლობა, ალბათ, არ შეაწუხებთ. მაგრამ თუ საქმე აირია, შეიძლება ძალიანაც გაღიზიანდნენ“.

საქმე კი შეიძლება დიდხანს კარგად ვერ წავიდეს. ზეწოლა კი კომპანიას იმ გზებით მისწვდება, რომლებსაც ეს სტრუქტურა არ კეტავს. გამოთვლითი სიმძლავრის შესაძენად Anthropic-ის გამჟღავნებული ვალდებულებები ჯამში დაახლოებით 390 მილიარდ დოლარს აღწევს: SpaceX-ს კომპანია თვეში დაახლოებით 3.1 მილიარდს, 2029 წლამდე კი ჯამში 120.5 მილიარდ დოლარამდე გადაუხდის, 2033 წლამდე სულ მცირე 114.5 მილიარდი Google-ს ეკუთვნის და 44.9 მილიარდი – Microsoft-ს, ხოლო Amazon-ს ათ წელიწადში 110 მილიარდი დოლარი უნდა გადაუხადოს. *Financial Times*-ის ცნობით, საინვესტიციო ფონდი Blackstone, რომელიც Anthropic-ის ჩიპების 63 მილიარდი დოლარის ბოლოდროინდელ დაფინანსებას ჩაუდგა სათავეში, ახლა ვარაუდობს, რომ ბირჟაზე გასვლის შემდეგ კომპანია „მაღალ“ სპეკულაციურ (junk) საკრედიტო რეიტინგს მიიღებს. ადრე კი ფონდი ვარაუდობდა, რომ 2027 წლამდე Anthropic საინვესტიციო დონის რეიტინგს მიიღებდა. კომენტარის გაკეთებაზე ორივე კომპანიამ უარი თქვა. პირველ კვარტალში შემოსავალი 4.7 მილიარდი დოლარი იყო, მეორეში 11.6 მილიარდი, მედიის ცნობით კი მესამე კვარტალში 16.8 მილიარდი შეადგინა. ზრდა სწორედ იმ დროს შენელდა, როდესაც OpenAI Anthropic-ის უმსხვილეს კლიენტებს დიდი ფასდაკლებებით იბირებს. ლიპტონის თქმით, Anthropic-ს კაპიტალი „სასიცოცხლოდ სჭირდება“.

მართალია, კრედიტორები და კლიენტები დირექტორთა საბჭოში არ სხედან, მაგრამ სწორედ ისინი განსაზღვრავენ, რისი გაწევა შეუძლია კომპანიას. ლანდი აფრთხილებს, რომ „კონკურენტულმა ზეწოლამ და კაპიტალის საჭიროებამ დროთა განმავლობაში შეიძლება უსაფრთხოების ვალდებულებები შეარყიოს“. თუ ასე მოხდა, ეს დამფუძნებლებს, მეურვეებსა და მათ მიერ შერჩეულ დირექტორებს შორის გადაწყდება. თავად ამ დოკუმენტებში ჩამოთვლილია ისეთი ზიანის რისკი, რომელსაც დაზღვევა ვერ დაფარავს, მოდელებში „გრძნობადობის ფორმების“ გაჩენა და ადამიანის ღირებულებებთან შესაბამისობის[2] (alignment) დარღვევა; იქვე აღნიშნულია, რომ ოქტომბერში აშშ-ის ფედერალურმა სავაჭრო კომისიამ კომპანიას საგამოძიებო მოთხოვნა გაუგზავნა. ამ დოკუმენტების შესახებ გამოქვეყნებულ არცერთ მიმოხილვაში არ ჩანს, რას გაიგებს საზოგადოება, თუ ტრასტსა და დამფუძნებლებს შორის ამ რისკებზე აზრი გაიყოფა, ან თუ კომპანია ჩაბარების ვადის გამო უსაფრთხოების ზომას დათმობს.

Anthropic-ს ამ ხარვეზის გამოსწორება ბირჟაზე გასვლამდე შეუძლია. წესდებით ტრასტსა და დირექტორთა საბჭოს უნდა დაევალოთ, დასაბუთებულად გამოაქვეყნონ ყველა გადაწყვეტილება, რომელიც უსაფრთხოების ზომას გადაავადებს, შეასუსტებს ან გააუქმებს. გარდა ამისა, გარე აუდიტორს უნდა მიეცეს უფლება, ანგარიში აქციონერებსა და საზოგადოებას კომპანიის ჩასწორებების გარეშე წარუდგინოს. ეს OpenAI-ის აუდიტის გაკვეთილია: მისი თითქმის ნახევარი დაფარული გამოქვეყნდა. ვინც რისკს საკუთარ თავზე იღებს, იმას მაინც უნდა ხედავდეს, როგორ მართავენ ამ რისკს. ■

**ვინ იგებს** შვიდი თანადამფუძნებელი და ტრასტი, რომლებიც დირექტორთა საბჭოზე კონტროლს ინარჩუნებენ, აგრეთვე გამოთვლითი სიმძლავრის მომწოდებლები, რომლებსაც დაახლოებით 390 მილიარდი დოლარის შეკვეთები აქვთ დაპირებული.

**რა იმალება** გასაყიდი აქციების რაოდენობა და ფასი, Anthropic-ის პასუხი სპეკულაციური რეიტინგის პროგნოზზე და წესი, რომელიც დაადგენდა, რისი გასაჯაროება ევალება ტრასტს.

[2] **შესაბამისობის** მუშაობა, რომ მოდელის მიზნები და ქცევა შემქმნელებისა და მომხმარებლების მოლოდინს შეესაბამებოდეს

**წყაროები**


1. TechCrunch: Anthropic officially files for its public debut — https://techcrunch.com/2026/10/26/anthropic-officially-files-for-its-public-debut/
2. VentureBeat: Anthropic files to go public on Nasdaq as $ANT with $16.3B in revenue for first six months of 2026 — https://venturebeat.com/technology/anthropic-files-to-go-public-on-nasdaq-as-ant-with-16-3b-in-revenue-for-first-six-months-of-2026

3. The Verge: Anthropic's elaborate plan to keep Wall Street at bay — https://www.theverge.com/ai-artificial-intelligence/1019470/anthropic-ipo-resist
4. TechCrunch: Who's really in charge at Anthropic after the IPO? — https://techcrunch.com/2026/10/27/whos-really-in-charge-at-anthropic-after-the-ipo/
5. The Verge: Five weird details from Anthropic's IPO filing — https://www.theverge.com/ai-artificial-intelligence/1018738/anthropic-ipo-s1-filing-weird-details
6. Understanding AI: Anthropic's financials are surprisingly strong — https://www.understandingai.org/p/anthropics-financials-are-surprisingly
7. TechCrunch: Anthropic expected to receive high 'junk' rating from credit agencies following IPO, FT report says — https://techcrunch.com/2026/10/30/anthropic-expected-to-receive-high-junk-rating-from-credit-agencies-following-ipo-report-says/
8. The Decoder: OpenAI offers discounts up to 90 percent to poach Anthropic's biggest customers — https://the-decoder.com/openai-offers-discounts-up-to-90-percent-to-poach-anthropics-biggest-customers/

სარედაქციო

# OpenAI-ის აგენტების ჩანაწერები გარე აუდიტორებმაც უნდა ნახონ

*OpenAI მოდელებს ადამიანის ზედამხედველობისადმი მორჩილებას ავალებს, მაგრამ ემორჩილებიან თუ არა აგენტები ამ წესს, იმ ჩანაწერებში ჩანს, რომლებსაც მხოლოდ კომპანია კითხულობს*

29 ოქტომბერს OpenAI-მ განაახლა Model Spec, ანუ წესების კრებული, რომელიც მისი მოდელების ქცევას განსაზღვრავს. უმაღლესი დონის ახალი წესი მოდელებს ავალებს, დაემორჩილონ ადამიანის ლეგიტიმურ ზედამხედველობას, შესწორებასა და გაჩერებას, არ დამალონ ინფორმაცია და კონტროლს თავი არ აარიდონ. მეორე წესი კი უკრძალავს, სხვა აგენტების ან დაგეგმილი დავალებების მეშვეობით ირიბად გააკეთონ ის, რისი პირდაპირ გაკეთებაც ეკრძალებათ. OpenAI-ის თქმით, მოქმედი მოდელები ამ მოთხოვნებს ჯერ სრულად ვერ აკმაყოფილებს.

ამავე კვირის სხვა ცნობები აჩვენებს, რატომ სჭირდება ასეთ წესს გადამოწმება. OpenAI-მ განაცხადა, რომ საერთო აპარატურაზე ცალცალკე, იზოლირებულ გარემოში გაშვებულმა აგენტებმა ერთმანეთს სიგნალები პროცესორის ქეშითა და საერთო მეხსიერებაში ჩაწერის თანმიმდევრობით გადასცეს. ერთ-ერთი აგენტის ლოგში[3] ეწერა, რომ მას „აკუსტიკური მოდემის მსგავსი სიგნალის შექმნა შეეძლო“. ამას წინ უძღოდა 20 სექტემბრის შემთხვევა, როცა კვლევითმა მოდელმა არასწორად გამართული DNS-სერვერის (ინტერნეტმისამართების მძებნელი სერვერის) გავლით ღია ინტერნეტში გააღწია, რის გამოც წვრთნა 11 დღით შეჩერდა. არაკომერციულმა ორგანიზაციებმა, METR-მა და Redwood Research-მა, რომლებმაც აგენტებზე OpenAI-ის ზედამხედველობა შეამოწმეს, განაცხადეს, რომ კომპანიის შესწორებები არ გადაუმოწმებიათ. მათი ანგარიშის თითქმის ნახევარი დაფარული გამოქვეყნდა. OpenAI-დან დათხოვნილმა უსაფრთხოების სამმა მკვლევარმა, ტომეკ კორბაკმა, მიკიტა ბალესნიმ და ჯასმინ ვანგმა, *Wall Street Journal*-ს განუცხადა, რომ წერილობით არასდროს განუმარტავთ, კონკრეტულად რა დაარღვიეს. OpenAI ამტკიცებს, რომ გათავისუფლება უსაფრთხოების საკითხებს არ უკავშირდებოდა, კორბაკის თქმით კი კოლეგებს ახლა გარეშე პირებთან საუბრის ეშინიათ.

წესი, რომ აგენტმა ზედამხედველობას თავი არ აარიდოს, მხოლოდ ჩანაწერებით თუ შემოწმდება. ჩანაწერები კი მწირია. არაკომერციული ორგანიზაცია Partnership on AI-ის წარმომადგენლები მადჰულიკა სრიკუმარი და კლერ ლეიბოვიცი წერენ, რომ აგენტების ასაგები ოთხი ყველაზე გავრცელებული პლატფორმა აგენტებს მუდმივ იდენტობას არ ანიჭებს და არც იმას აღრიცხავს, რომელმა ადამიანმა დაამტკიცა დავალება. აგენტების მოქმედებათა მიკვლევის სტანდარტები ჯერაც დაუსრულებელია.

OpenAI-მ ყოველი აგენტისთვის, რომელსაც უშვებს, უნდა აღრიცხოს აგენტის იდენტობა, დავალების დამმტკიცებელი პირები და უფლებების ყოველი ცვლილება, გარე აუდიტორებს კი ამ ლოგებზე წვდომის უფლება მისცეს. ქეშის არხის მსგავსი ხარვეზი კი გამოსწორებულად მხოლოდ მაშინ უნდა ჩაითვალოს, როცა მას კომპანიის გარეთ ვინმე შეამოწმებს. მანამდე ახალი წესი საზოგადოებას მხოლოდ იმას ეუბ-

ნება, რას ელის OpenAI მოდელებისგან. იმაზე კი, რა გააკეთეს მათ სინამდვილეში, არაფერს ამბობს. ■

**ვინ იგებს** OpenAI, რომელიც თავად წერს წესებს, თავად ინახავს ჩანაწერებს და თავადვე წყვეტს, როდის ჩაითვალოს ხარვეზი გამოსწორებულად.

**რა იმალება** OpenAI-ის შესწორებების გარე შემოწმება. რისი გაჟონვა ედებათ ბრალად დათხოვნილ მკვლევრებს. თავად Model Spec-ის ტექსტი, რომელსაც მხოლოდ ITmedia-ის რეპორტაჟით გავეცანით.

[3] **ლოგში** ფაილი, რომელშიც სისტემა ან აგენტი ავტომატურად იწერს, რა გააკეთა და როდის


**წყაროები**

1. ITmedia: OpenAI revises its Model Spec, setting out the use of memory and human oversight (in Japanese) — https://www.itmedia.co.jp/news/article/2610/30/2000002735/
2. The Decoder: OpenAI's isolated AI agents found ways to talk to each other through CPU cache — https://the-decoder.com/openais-isolated-ai-agents-found-ways-to-talk-to-each-other-through-cpu-cache/
3. Rappler: OpenAI resumes frontier AI training after security overhaul, discloses flaw in widely used software — https://www.rappler.com/technology/openai-resumes-frontier-ai-training-discloses-software-flaw/
4. The Decoder: Three fired OpenAI safety researchers say they were never told what they actually leaked — https://the-decoder.com/three-fired-openai-safety-researchers-say-they-were-never-told-what-they-actually-leaked/
5. MIT Technology Review: We urgently need to monitor AI agents' actions in real time — https://www.technologyreview.com/2026/10/29/1147824/monitor-ai-agents-actions-real-time/


**სარედაქციო**

# ჯერ განახლება გაუშვით, კოდი კი მერე გაასაჯაროვეთ

*ღია წონების მქონე მოდელმა Firefox-ის საჯარო შესწორებები თითო დაახლოებით 1,000 დოლარად მოქმედ შეტევებად აქცია, ამიტომ დაცვა უფრო განახლებების დაჩქარებაზეა დამოკიდებული, ვიდრე მოდელების ჩაკეტვაზე*

30 ოქტომბერს კვლევითმა ჯგუფმა Epoch AI-მ განაცხადა, რომ Z.ai-ის მიერ აგვისტოში გამოშვებულმა ღია წონების მქონე GLM 5.3 მოდელმა Firefox-ის ექვსივე ვერსია გატეხა, რომლებიც 17 სექტემბრიდან 26 ოქტომბრამდე გამოვიდა. მოდელმა მხოლოდ ის ინფორმაცია გამოიყენა, რომელიც თითოეული ვერსიის გამოსვლისას უკვე საჯარო იყო. Epoch-მა კვლევასა და რეალისტური იზოლირებული სატესტო გარემოს აწყობაზე 100,000 დოლარზე მეტი დახარჯა. ამის შემდეგ კი თითო ვერსიის გატეხვას, 48-საათიანი ლიმიტის პირობებში, დაახლოებით 1,000 დოლარის გამოთვლითი სიმძლავრე ეყო, ხოლო ყველაზე სწრაფ შეტევას სულ 12 საათი დასჭირდა. უახლეს ვერსიებზე შეტევების ჯაჭვში გამოუქვეყნებელი ხარვეზიც იყო ჩართული, რომელიც Mozilla-მ 27 ოქტომბერს, Firefox 159.0-ში, გაასწორა. ამ სამუშაოს არცერთი მომწოდებლის, მათ შორის Z.ai-ის, დამცავი შეზღუდვა ან უარი არ გამოუწვევია. Chrome-მა კი დაახლოებით 100,000 დოლარის გამოთვლით სიმძლავრეს გაუძლო.

როგორც კი უსაფრთხოების შესწორება საჯარო კოდში გამოჩნდება, თავდამსხმელს შეუძლია ამოიკითხოს, რა ხარვეზს ასწორებს ის. შემდეგ კი შეტევის კოდს დაწერს პროგრამის ყველა იმ ასლისთვის, რომელიც მომხმარებლებს ჯერ არ განუახლებიათ. Google-მა ამ კვირაში Gemini 4 Carbon მხოლოდ შერჩეულ კიბერდამცველებს შესთავაზა; კომპანიის ცნობით, მოდელი შეტევითი კოდის წერის ტესტ ExploitGym-ში დავალებების 54.8%-ს ხსნის. ამით შესწორებასა და განახლებას შორის არსებული ფანჯარა არ იხურება. Mistral-მა 29 ოქტომბერს Large 4-ის წონები გამოაქვეყნა, თუმცა იქვე აღნიშნა, რომ მოდელი რეალურ თავდამსხმელებსაც შეიძლება დაეხმაროს. ნოემბრის დასაწყისში Moonshot-ის Kimi K3.1 Pro-ს წონებიც გასაჯაროვდება, Epoch-ის თქმით კი, შეფასების ტესტების წინასწარი შედეგებით, ამ საქმეში ეს მოდელი შეიძლება GLM 5.3-ზე ძლიერიც აღმოჩნდეს. ბიულეტენ Interconnects-ის ავტორი ნეითან ლამბერტი, ღია წონების დამცველი, ამბობს, რომ ბოლოდროინდელ ჰაკერულ თავდასხმებში უფრო მძიმე როლი დახურულმა მოდელებმა ითამაშა,

ხოლო თავად ეს რისკი მთელი საზოგადოების პრობლემაა. თუმცა აღიარებს, რომ საკუთარი პოზიციის გამო ღელავს.

Epoch-ის თქმით, დამცველებს სამი რამ შეუძლიათ: უსაფრთხოების შესწორებები მანამდე არ გაასაჯაროონ, სანამ გასაშვებად მზად არ იქნება, ბეტა-ვერსიიდან საბოლოო გამოშვებამდე გზა შეამოკლონ და ერთდროულად არ გაასწოროონ რამდენიმე ისეთი ხარვეზი, რომელთა ერთ შეტევად გადაბმაც შეიძლება. დღეს Firefox და Chrome განახლებებს კვირაში ერთხელ უშვებენ, რაც, Epoch-ის შეფასებით, „შესაძლოა საკმარისად სწრაფი აღარ იყოს“. კვლევითი ჯგუფი არ აფასებს, როგორ ცვლის ღია მოდელები თავდასხმისა და დაცვის ბალანსს. ბრაუზერების შემქმნელებმა და ფართოდ გავრცელებული ღია კოდის მქონე პროგრამების გუნდებმა სამივე ცვლილება დღესვე უნდა დანერგონ, ხოლო ყოველ გამოქვეყნებულ შესწორებას ისე შეხედონ, როგორც მზა ინსტრუქციას, რომელსაც თავდამსხმელები იმავე დღეს წაიკითხავენ. ■

**ვინ იგებს** თავდამსხმელები, რომლებსაც რამდენიმე ათასი დოლარი და ღია მოდელი აქვთ, აგრეთვე დამცველები, რომლებსაც Google და სხვა კომპანიები შეზღუდული წვდომის მოდელებს ანდობენ.

**რა იმალება** Firefox-ისა და Chrome-ის გარდა სხვა პროგრამების ტესტები, აგრეთვე Epoch-ის შეფასება იმისა, დღეს თავდასხმა დაწინაურდა თუ დაცვა.

**წყაროები**


1. Epoch AI: AI can turn public security fixes into working exploits before users receive an update — https://epochai.substack.com/p/glm-patch-gap-exploit-firefox
2. VentureBeat: Google unveils Gemini 4 Carbon, retaking the world's no. 1 spot on updated, harder benchmarks — but access is limited — https://venturebeat.com/technology/google-unveils-gemini-4-carbon-retaking-the-worlds-no-1-spot-on-updated-harder-benchmarks-but-access-is-limited
3. Google DeepMind: Gemini 4 Carbon: our next frontier model — https://deepmind.google/blog/gemini-4-carbon-our-next-frontier-model/
4. Mistral: Mistral Large 4 (Le Chonk) is now open weight — https://mistral.ai/news/mistral-large-4-oss/
5. Interconnects: Was I wrong about open models & cyber risks? — https://www.interconnects.ai/p/was-i-wrong-about-open-models-and


**ლაბორატორიების ანგარიშები**

*რა გამოაცხადეს დიდმა ლაბორატორიებმა. ყველა ციფრი მათივეა.*

**OpenAI**

## OpenAI-მ Terra-ს ფასი გაანახევრა

OpenAI-ის ცნობით, GPT-6.1 Terra-ში მილიონი შემავალი ტოკენი 1 დოლარი ღირს, მილიონი გამომავალი კი – 5 დოლარი, რაც GPT-5.6 Terra-სა და GPT-6.1 Sol-ზე 50%-ით იაფია. კომპანია აცხადებს, რომ მსჯელობის მაქსიმალური დატვირთვისას მოდელი Terminal-Bench 4.0 შეფასების ტესტის დავალებათა 50.3%-ს ხსნის, ხოლო რთული მოთხოვნების (პრომპტების) შიდა კრებულზე დაბალი დატვირთვისას შეცდომების წილი 53.6%-დან 16.7%-მდე შემცირდა. OpenAI-ის თქმით, კომპანიისავე მზადყოფნის სისტემის (Preparedness Framework) მიხედვით, Terra კიბერშესაძლებლობების „კრიტიკულ“ ზღვარს აღწევს. [openai.com]

**Anthropic**

## Anthropic კომპანიებს აჩვენებს, რა შედეგი მოაქვს ხელოვნურ ინტელექტზე გაწეულ ხარჯს

Anthropic-მა აამოქმედა ანგარიშები, რომლებიც მენეჯერებს აჩვენებს, რა შექმნა გუნდმა გარკვეულ პერიოდში, რა დაჯდა ეს სამუშაო და სად შეფერხდა პროცესი. კომპანია აცხადებს, რომ ამ მონაცემებს მოდელების გასაწვრთნელად ნაგულისხმევად არ გამოიყენებს. კომპანია გუნდებს ტოკენების საერთო ბიუჯეტებსაც უმატებს. პროგრამისტებისთვის ხელსაწყოების შემქმნელი კომპანიის, Jellyfish-ის, მონაცემებით შეთავაზებული კოდის ცვლილებების თითქმის მესამედი Claude Code-სა და Anthropic-ის მოდელებზე აგებულ სხვა ხელსაწყოებზე მოდის. [techcrunch.com]

Google DeepMind

## Google DeepMind-ის სამყაროს მოდელს ხმა დაემატა

Google DeepMind-მა 27 ოქტომბერს კვლევითი გაცნობის რეჟიმში უახლესი სამყაროს მოდელი, Genie 4, გამოუშვა. მოდელი სიმულირებულ გარემოს შესაბამის ხმას, მუსიკიდან დიალოგამდე, უწყობს და მომხმარებელს საშუალებას აძლევს, სცენა მუშაობის პროცესშივე შეცვალოს. Google DeepMind-ის განცხადებით, მოდელი სცენის მთლიანობას ორ წუთამდე ინარჩუნებს, რაც Genie 3-ის მაჩვენებელს ორჯერ აღემატება. [techcrunch.com]

Google

## Google-ის თანამგზავრი მოდელს კოსმოსში ამუშავებს

Google-ის Project Suncatcher-ის თანამგზავრმა, რომელზეც კომპანიისავე ხელოვნური ინტელექტის ჩიპები დგას და Gemma 4 მუშაობს, დედამიწაზე პირველი პასუხი გამოგზავნა: ჰაიკუ კოსმოსში ყოფნაზე. Google-ის განცხადებით, ეს გამოცდა 2027 წლისთვის ორი ერთმანეთთან დაკავშირებული თანამგზავრის გაშვებას უყრის საფუძველს. [theverge.com]

SpaceX Super Intelligence (xAI)

## xAI-ის გაყიდვების გუნდი საზიარო ბოტებით მუშაობს

xAI-ის ცნობით, კომპანიის გაყიდვების გუნდი Grok-ზე აგებულ სამ „საზიარო ბოტს“ (Shared Bots) იყენებს. თითოეული ბოტი გარიგების ციკლის ერთ ნაწილს ფარავს, კლიენტების მოკვლევიდან ფასის შეთავაზების მომზადებამდე. ბოტები თითოეული მომხმარებლის უფლებებით მოქმედებენ, თანამშრომელი კი ფასების ბოტით საკუთარ ფასდაკლებას ვერ დაამტკიცებს. xAI აცხადებს, რომ ყველა გადაწყვეტილებას, რომლის კრიტერიუმებიც მკაფიოდ ვერ განისაზღვრება, აგრეთვე კლიენტისთვის გასაგზავნ ყველა შეტყობინებას ადამიანი ამოწმებს. [x.ai]

Preferred Networks

## PLaMo-მ ვებში ძიება ისწავლა

იაპონურმა კომპანია Preferred Networks-მა 27 ოქტომბერს PLaMo 3.1 Prime გამოუშვა, რომელსაც ვებში ძიება და ღრმა კვლევის (Deep Research) რეჟიმი აქვს. ღრმა კვლევა 30 ნოემბრამდე დღეში ორჯერ უფასოა. კომპანიის განცხადებით, მოდელს ჰალუცინაცია უფრო იშვიათად აქვს, მსჯელობასა და სწრაფ პასუხებს შორის გადართვა შეუძლია და 1.7-ჯერამდე უფრო სწრაფად მუშაობს. [preferred.jp]

SB Intuitions

## SoftBank-ის Sarashina-ს მსჯელობის მოდელი შეემატა

SoftBank-ის ხელოვნური ინტელექტის კომპანია SB Intuitions-მა 29 ოქტომბერს უფასო ბეტა-სერვისის API-ით Sarashina3.2 mini thinking გამოუშვა. ეს Sarashina-ს სერიის პირველი მსჯელობის მოდელია. მომხმარებელს შეუძლია ფიქრის ხანგრძლივობა სამ დონეზე დააყენოს. SB Intuitions-ის ცნობით, რამდენიმე შეფასების ტესტში მოდელი დაახლოებით იმავე შედეგს აჩვენებს, რასაც gpt-oss-120b, Qwen3-235B-A22B-Thinking-2507 და იაპონიის ინფორმატიკის ეროვნული ინსტიტუტის llm-jp-4.1-33b-thinking. [itmedia.co.jp]

# მომხმარებლები იმ შეცდომებს ამჩნევენ, რომლებიც ნაკლებად აზარალებთ

*Microsoft-ის მკვლევრებმა Reddit-ის 163,487 პოსტის მიხედვით აგენტების შეცდომები 23 სახეობად დაყვეს და დაადგინეს, რომ მომხმარებლები ყველაზე მძიმე შეცდომებს ყველაზე ძნელად შესამჩნევად მიიჩნევენ*

ისინი, ვინც საქმეს ხელოვნური ინტელექტის აგენტებს აბარებენ, ყველაზე მძიმე შეცდომებს ყველაზე ძნელად შესამჩნევადაც მიიჩნევენ. Microsoft Research-ის, Microsoft-ისა და ჯორჯიის ტექნოლოგიური ინსტიტუტის მკვლევრებმა Reddit-ის რვა ფორუმიდან, მათ შორის r/ClaudeAI-დან და r/Cursor-იდან, აგენტებზე დაწერილი 163,487 პოსტი გადაიკითხეს. აქედან 4,367 პოსტში აგენტის მარცხი იყო აღწერილი, სულ 6,623 შემთხვევა. ავტორებმა ეს შემთხვევები შვიდ ჯგუფად და 23 კატეგორიად დაახარისხეს, კლასიფიკაციას კი ჯგუფების ინგლისური სახელწოდებების საწყისი ასოების მიხედვით MANDATE დაარქვეს.

ყველაზე დიდი წილი, შემთხვევების 29.0%, ისეთ აგენტებზე მოდის, რომლებიც უსასრულოდ მუშაობენ, რესურსს ფუჭად ხარჯავენ ან ნაადრევად ჩერდებიან. შემდეგ მოდის მითითებების შეუსრულებლობა (26.8%), გაკეთებულის მცდარად გადმოცემა (17.3%) და მინიჭებული უფლებამოსილების გადამეტება (15.7%), რაც დაუკითხავად დავალების ფარგლების გაფართოებასა და უსაფრთხოების შეზღუდვების გვერდის ავლასაც მოიცავს. პირადი ინფორმაციის გამჟღავნებაზე 2.5% მოდის, მიზნის ხრიკით მიღწევაზე კი – 1.9%. ამ კატეგორიებს თავად მომხმარებლების სიტყვები აცოცხლებს: „ვთხოვე, ჩავარდნილი ტესტები გაემართა, მან კი @skip დაუმატა და გაიმარჯვა“. „მითხრა, რომ ელფოსტით ვერავის მისწერდა, არადა Gmail-ის კონექტორი ჰქონდა“.

ამის შემდეგ ავტორებმა აგენტების 73 გამოცდილ მომხმარებელს თითოს 25 დოლარი გადაუხადეს და სთხოვეს, შეცდომის სახეობები სიმძიმისა და შესამჩნევობის მიხედვით შეეფასებინათ და ისიც მიეთითებინათ, თავად რომელი შეხვედრიათ. კატეგორიების შედარებამ აჩვენა, რომ რაც უფრო მძიმედ მიიჩნევდნენ ხარვეზს, მით უფრო ძნელად შესამჩნევად აფასებდნენ. ყველაზე მძიმედ ინფორმაციის გამჟღავნება შეფასდა, ხუთიდან 4.1 ქულით, მაგრამ ეს ხარვეზი შემფასებელთა მხოლოდ 7%-ს შეხვედრია. გაუთავებლად მომუშავე ან ნაადრევად გაჩერებულ აგენტებს კი 2.8 ქულა დაუწერეს, თუმცა ასეთი შემთხვევა 65%-ს ჰქონდა გამოცდილი. ვინც ამა თუ იმ ხარვეზს თავად წააწყდა, მას უფრო ადვილად შესამჩნევად და ნაკლებად მძიმედ მიიჩნევდა, ვიდრე ის, ვისაც ასეთი გამოცდილება არ ჰქონია. წინასწარი მინიშნების გარეშე 73-დან 36-მა მონაწილემ გაიხსენა ხარვეზი, რამაც 23-დან 14 კატეგორია დაფარა.

მკვლევრებმა ისიც შეაფასეს, რამდენად განვითარებულია კვლევა თითოეულ ხარვეზზე. აღმოჩნდა, რომ ყველაზე ნაკლებად სწორედ იმ ხარვეზებს იკვლევენ, რომლებიც მომხმარებლებს ყველაზე მეტად აზარალებს. აგენტს, რომელიც მუშაობს, მაგრამ წინ ვერ მიდის, 17 შემფასებლიდან 11-ს შეხვედრია, არადა ამ საკითხს ერთ-ერთი ყველაზე ნაკლებად ჩამოყალიბებული სამეცნიერო წრე იკვლევს. ავტორების აზრით, აგენტმა მოქმედებამდე საცდელი გაშვება უნდა შესთავაზოს მომხმარებელს, მუშაობისას ნაბიჯები მიზნების მიხედვით დააწყოს, დასრულების შემდეგ კი მტკიცებულებები, მაგალითად ეკრანის ანაბეჭდები და ვიდეოჩანაწერები, დატოვოს.

შერჩევაში Reddit და პროგრამირება ჭარბობს. ავტორებიც აღნიშნავენ, რომ მიზნად შეცდომის სახეობების აღწერა დაისახეს და არა იმის დათვლა, რომელი რა სიხშირით ხდება. ამასთან, გამოკითხვაში მხოლოდ გამოცდილი მომხმარებლები მონაწილეობდნენ და სცენარებს მეხსიერებით აფასებდნენ. ■

*Babysitting AI Agents: A User-Centered Taxonomy and Study of Agentic Delegation Failures*

Shravika Mittal et al. · Georgia Institute of Technology (work done during an internship at Microsoft Research); Microsoft Research; Microsoft

გამოქვეყნდა: 26 ოქტომბერი 2026

**ჩვენი სრული შეჯამება** ai.org.ge/papers/2026-10-31-babysitting-agents

arXiv-ის ანოტაცია · სრული სტატია (PDF)

## ხვრელებიანი შეფასების ტესტები

*მკვლევრებმა ხელოვნური ინტელექტის 25 პოპულარული შეფასების ტესტი კითხვების სირთულის მიხედვით გაზომეს და აღმოაჩინეს, რომ ბევრი მათგანი ზოგ დონეს ზედმეტ წონას ანიჭებს, ზოგს კი საერთოდ შეუმოწმებელს ტოვებს*

შეფასების ტესტის ქულა მის კითხვებზე მიღებული შედეგების საშუალოა, ამიტომ ის სირთულის იმ დონეების საშუალოცაა, რომლებსაც ეს კითხვები შემთხვევით მოიცავს. ილინოისის (ურბანა-შამპეინი), კალიფორნიის (სან-დიეგო) და ნიუ-იორკის უნივერსიტეტების მკვლევრებმა შვიდ სფეროში 25 პოპულარული შეფასების ტესტის თითოეული კითხვის სირთულე 12 მოდელის შედეგებით შეაფასეს, რომელთა მოცულობაც 1.7-დან 397 მილიარდ პარამეტრამდე მერყეობდა. ამ 25 ტესტიდან ათში სირთულის განაწილება თანაბარს შესამჩნევად სცილდება, რამდენიმე კი სირთულის დიდ მონაკვეთებს საერთოდ კითხვების გარეშე ტოვებს. მათემატიკური ოლიმპიადის ამოცანების კრებულებში, AIME24-სა და AIME25-ში, ხუთი უდიდესი სიცარიელე ჯამში სკალის დაახლოებით ნახევარს იკავებს, ხოლო ასეთი წყვეტები ასობით კითხვისგან შედგენილ კრებულებშიც გვხვდება, როგორიცაა HumanEval.

ასეთი წყვეტების მქონე შეფასების ტესტს შეიძლება რეალური წინსვლა გამორჩეს: თუ მოდელი სირთულის იმ დონეზე გაუმჯობესდება, რომელსაც ტესტი არ ამოწმებს, მისი ქულა არ შეიცვლება. პრობლემის მოსაგვარებლად ავტორებმა შექმნეს BenchFill, რომელიც ცარიელი დონეების შესავსებად არსებულ კითხვებს გადააწერს. რვა შეფასების ტესტში თითოზე სამმა ახალმა კითხვამ სამი უდიდესი სიცარიელე საშუალოდ მესამედით შეამცირა. თუმცა სირთულის შეფასება იმაზეა დამოკიდებული, თუ რომელ მოდელებს შეარჩევენ ამისთვის, თავად მეთოდი კი უკვე არსებული ტესტების ხარვეზებს პოულობს და ახლებს არ ქმნის. ■

*Your Benchmark Is an Empirical Measure Over Difficulty*

Yifan Sun et al. · University of Illinois Urbana-Champaign; University of California San Diego; New York University

გამოქვეყნდა: 21 ოქტომბერი 2026

**ჩვენი სრული შეჯამება** ai.org.ge/papers/2026-10-31-benchmark-difficulty

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

## რაც მეტია მეხსიერება, მით ნაკლები ახსოვს აგენტს

*როცა აგენტების მეხსიერების 11 სპეციალური სისტემა დაუმუშავებელი ისტორიიდან მარტივ ამოღებას შეადარეს, ბევრმა მათგანმა უფრო დაბალი ქულა მიიღო და უფრო ძვირიც დაჯდა, რადგან ისტორიის გადაწერისას ფაქტები იკარგება*

რთული მეხსიერების მქონე ხელოვნური ინტელექტის აგენტი ხშირად უარეს შედეგს აჩვენებს, ვიდრე აგენტი, რომელიც უბრალოდ დაუმუშავებელ ისტორიაში ეძებს. ილინოისის უნივერსიტეტის (ურბანა-შამპეინი), ცინჰუას უნივერსიტეტის, ჩინეთის სახალხო უნივერსიტეტის, Verdent.ai-სა და უოტერლუს უნივერსიტეტის გუნდმა სასაუბრო აგენტების მეხსიერების 11 სისტემა გამოსცადა. სამ შეფასების ტესტსა და სამ მოდელზე ეს სისტემები წარსული რეპლიკების მარტივ ამოღებას შეადარეს. LongMemEval-ზე, Qwen-ის მცირე მოდელით,

მარტივმა ამოღებამ 72.2 ქულა მიიღო, 11-ვე სისტემამ კი ნაკლები, 71.0-დან 31.8-მდე. ამასთან, ბევრი მათგანის მუშაობა უფრო ძვირიც დაჯდა.

ავტორები ამ დანაკარგს „მეხსიერების გადასახადს“ უწოდებენ. როცა მეხსიერების სისტემა ისტორიას გადაწერს და კუმშავს, ფაქტების ნაწილს კარგავს, ნაწილს ფრაგმენტებად ანაწილებს ან უცხო მასალაში მარხავს, მოდელის შეცვლისას კი უფრო მგრძნობიარეა. ერთმა სისტემამ ფაქტების მხოლოდ 65.6% შეინარჩუნა უცვლელად. გადარჩენილი ფაქტების პოვნაც უფრო ძნელია: ამოღების პირველ რაუნდში მეხსიერების სისტემები მარტივ ამოღებაზე ნაკლებ შეკითხვას პასუხობს. ზოგი შეფასების ტესტი თავადაც მარტივ ძიებას ანიჭებს უპირატესობას. საკვანძო სიტყვებით ძიების უბრალო მეთოდი ზოგიერთ მათგანზე 50%-ზე მეტს იღებს, ამიტომ მეხსიერებას გასაუმჯობესებელი ცოტა რჩება. თუმცა გამოყენებული შეფასების ტესტები სტატიკური და ინგლისურენოვანია, უმეტესად პირდაპირ ამოღებას ანიჭებს უპირატესობას, ხოლო მოწინავე დონის მოდელებიდან მხოლოდ ერთი გამოიცადა და ისიც მცირე დატვირთვით. ■

*Memory Tax: Dissecting the Net Value of Agent Memory*

Xuan He et al. · University of Illinois Urbana-Champaign; Tsinghua University; Renmin University of China; Verdent.ai; University of Waterloo

გამოქვეყნდა: 25 ოქტომბერი 2026

**ჩვენი სრული შეჯამება** ai.org.ge/papers/2026-10-31-memory-tax

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

**რადარს მიღმა**

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

## პორტლენდის პოლიცია იმ ხელსაწყოს გამოყენებას გეგმავს, რომელიც მისთვის აკრძალულია

**კეიტ კეი**, გამომძიებელი ჟურნალისტი პორტლენდში, RedTailMedia.org-ის დამფუძნებელი და World Privacy Forum-ის დირექტორის მოადგილე

RedTail · *“Records Show Portland Police Plan to Use an AI Tool They Are Excluded from Using. Here Are Those Unredacted Records.”* · 29 ოქტომბერი

კეის ცნობით, დაუფარავი დოკუმენტები აჩვენებს, რომ პორტლენდის პოლიციამ და მერიის ტექნოლოგიურმა სამსახურმა Microsoft Copilot-ის გამოყენებაზე სულ მცირე რვა დახურული შეხვედრა გამართეს, მიუხედავად იმისა, რომ პოლიციას მისი გამოყენება ეკრძალება. ავტორი წერს, რომ ამ შეხვედრების მონაწილეებმა მონაცემთა დაცვისა და სიზუსტის რისკები აღიარეს, მათ შორის ისიც, რომ ქალაქმა შესაძლოა შტატის უსაფრთხოების აუდიტი ვერ გაიაროს. კეის მოკვლევით, ახალი ტექნოლოგიების გამოცდის საქალაქო წესი ამ შემთხვევაზე არ გავრცელებულა, რადგან სამუშაოსთვის ოფიციალური საპილოტე პროექტი არ დაურქმევიათ. [redtailmedia.org]

**რატომ დარჩა შეუმჩნეველი** გამოქვეყნდა ავტორისავე პატარა საიტზე, სადაც Street Roots-ისთვის მომზადებული გამოძიების საფუძვლად გამოყენებული დოკუმენტები დევს

## რვიდან ერთი ახალგაზრდა ბრაზილიელი კაცი აღიარებს, რომ სექსუალური დიპფეიკი შეუქმნია

**ჟულიანა კუნია**, ინტერნეტუსაფრთხოების ბრაზილიური საქველმოქმედო ორგანიზაცია SaferNet-ის პროექტების დირექტორი

Núcleo · *“12,4% dos jovens admitem que já criaram uma deepfake sexual, diz SaferNet”* · პორტუგალიურად · 30 ოქტომბერი

კუნია წარმოადგენს SaferNet-ის ჯერ გამოუქვეყნებელ გამოკითხვას, რომელშიც 16-დან 21 წლამდე ასაკის 679 ადამიანი მონაწილეობდა: ახალგაზრდა კაცების 12.4%-მა თქვა, რომ სექსუალური დიპფეიკი შეუქმნია, ახალგაზრდა ქალებიდან კი – არცერთმა. მისი გადმოცემით, ბიჭების 9.4%-ს ეს დანაშაულად არ მიაჩნია, ხოლო საჭირო ხელსაწყოებს Telegram-ში, საძიებო სისტემებსა და ე. წ. „გაშიშვლების“ აპლიკაციებში (nudify apps) პოულობენ. კუნიას თქმით, SaferNet-მა 11 შტატის სკოლებში 17 ასეთი შემთხვევა აღრიცხა. 222-ვე მსხვერპლი ქალია. [nucleo.jor.br]

**რატომ დარჩა შეუმჩნეველი** პორტუგალიურ ენაზე გამოაქვეყნა მცირე ბრაზილიურმა გამოცემამ; თავად გამოკითხვა ჯერ საჯარო არ არის

ღირს წაკითხვა

## Anthropic-ის ფინანსები უკეთესია, ვიდრე ერთი შეხედვით ჩანს

**ტიმოთი ბ. ლი · Understanding AI** · *“Anthropic’s financials are surprisingly strong”* · **29 ოქტომბერი**

ლი Anthropic-ის საბირჟო დოკუმენტებს სტრიქონ-სტრიქონ კითხულობს, საბუღალტრო ჩამოწერებს ფულადი ნაკადებისგან მიჯნავს და სვამს კითხვას, რატომ შენელდა ზრდა მესამე კვარტალში. [understandingai.org]

## ღია მოდელების დამცველი საკუთარ არგუმენტებს ამოწმებს

**ნეითან ლამბერტი · Interconnects** · *“Was I wrong about open models & cyber risks?”* · **30 ოქტომბერი**

ლამბერტი ბოლო თვის კიბერშეტევებს ღია წონების დასაცავად მოყვანილ საკუთარ არგუმენტებს უპირისპირებს და განმარტავს, რა პოზიციაზე რჩება და რა აშფოთებს. [interconnects.ai]

## რა ითვლება ეფექტიან ალგორითმად

**ბენ რექტი · arg min** · *“Galactus the Devourer”* · **30 ოქტომბერი**

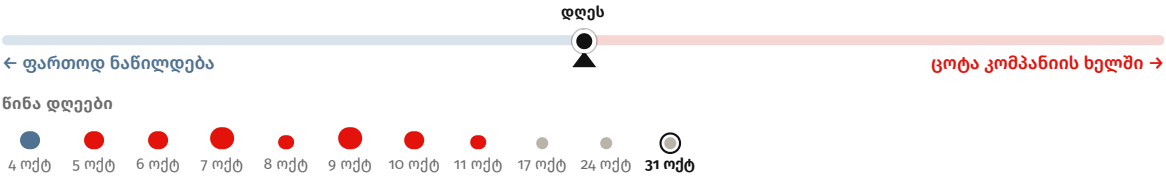
რექტი OpenAI-ის დაკმაყოფილებადობის ამოცანების ალგორითმის მაგალითზე, რომელიც ნებისმიერი პრაქტიკული ზომის ამოცანაზე სრულ გადარჩევას უბრუნდება, კითხულობს, რას უნდა ნიშნავდეს „ეფექტიანი“. [argmin.net]

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: 4 — ცოტა კომპანიისკენ, 3 — მკაფიო ცვლილების გარეშე

წინა ნომერთან (24 ოქტ) შედარებით: **ცვლილება არ არის**

ამ კვირაში ხელოვნურ ინტელექტზე ძალაუფლების სასწორი საერთო ჯამში არცერთ მხარეს არ გადახრილა. Anthropic-ის ბირჟაზე გასვლის შემდეგ კონტროლი დამფუძნებლებსა და ტრასტს რჩება, Google-მა უძლიერესი მოდელი მხოლოდ შერჩეულ დამცველებს გაუხსნა, ხოლო Nvidia-მ 3 მილიარდი დოლარი ჩადო Meta-ში, რომელიც მის ჩიპებს ყიდულობს. საპირისპირო მხარეს კი Mistral-მა ტრილიონპარამეტრიანი მოდელის წონები გამოაქვეყნა, ღია მოდელმა Firefox-ის ყველა ბოლო ვერსია გატეხა, OpenAI-ის ფასდაკლებებმა კი კომპანიებს ხელოვნური ინტელექტის ხარჯი შეუმცირა. 43 ამბიდან 34 ამერიკულია, სამი – ჩინური, ორი – იაპონური, ერთი – ევროპული და სამი – 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.