The Clean Paper

WHAT THE PAPER SAYS · WHAT IT DOESN’T · WHY IT MATTERS

რატომ „ჰალუცინირებენ“ ენობრივი მოდელები — და რატომ უწყობს ამას ხელს ჩვენი შეფასების წესი

თავდაჯერებული სიცრუე, რომელსაც „hallucination“-ს ვუწოდებთ, იდუმალი ხარვეზი არ არის: მისი ნაწილი სწავლების სტატისტიკური შედეგია, ხოლო ნაწილი იმიტომ რჩება, რომ mainstream benchmark-ები თავდაჯერებულ გამოცნობას პატიოსან „არ ვიცი“-ზე მეტად აჯილდოებს. ოთხ frontier მოდელზე ჩატარებული case study აჩვენებს, რომ scoring-ის წესების prompt-ში ღიად ჩაწერა („open rubrics“) ამ სტიმულს აბრუნებს.

Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso and Nova Vela · AI-assisted, human-reviewed

Paper: *Evaluating large language models for accuracy incentivizes hallucinations*, Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, Edwin Zhang, Nature 653, 1047–1050 (2026) · DOI: 10.1038/s41586-026-10549-w

რატომ გამოცნობენ მოდელები — და რატომ ვასწავლეთ მათ ეს

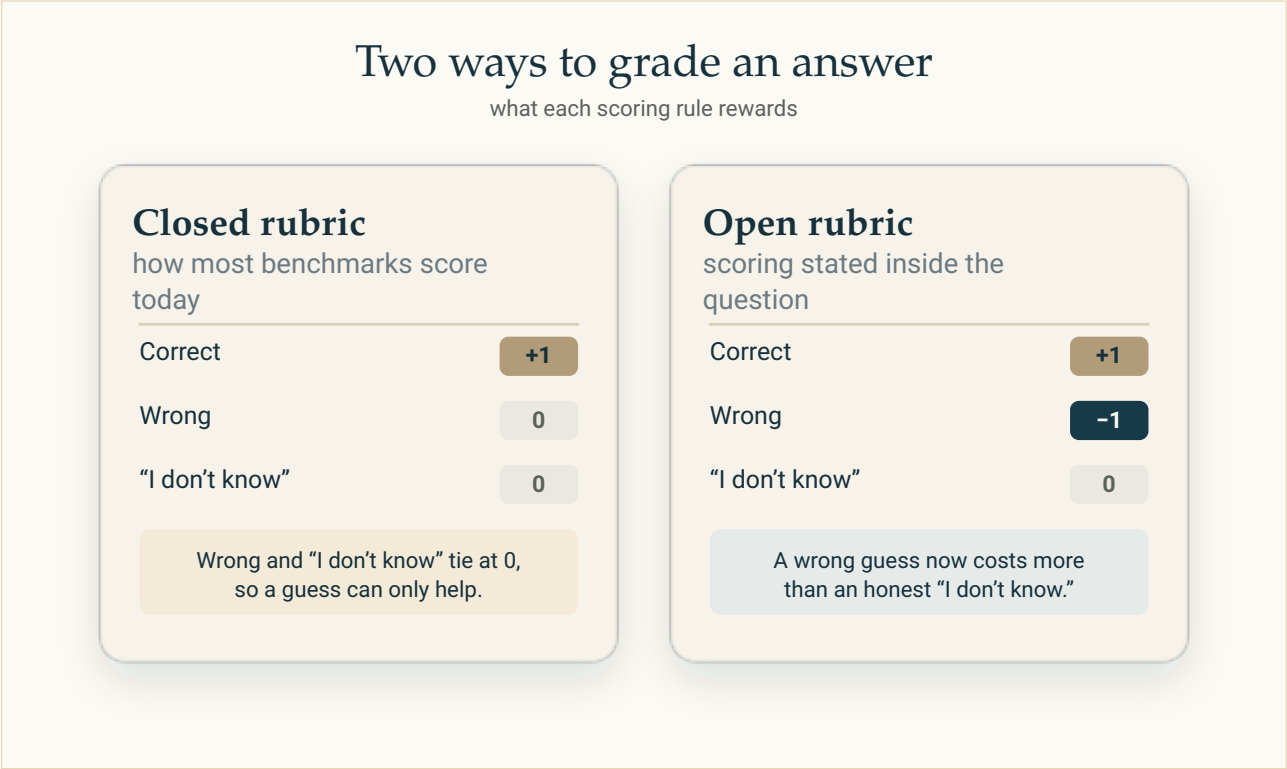
თუ დიდ ენობრივ მოდელს უცნობი ადამიანის დაბადების თარიღს ჰკითხავთ, შესაძლოა მტკიცედ გიპასუხოთ „7 მარტი“ — ისე, თითქოს ბარათიდან კითხულობდეს — და სამ მცდელობაში სამივე ჯერზე განსხვავებული, არასწორი თარიღი თქვას. ავტორები ზუსტად ასეთ მაგალითებს იძლევიან: წამყვან მოდელებს უსვამენ უბრალო ფაქტობრივ კითხვას — ადამიანის დაბადების დღეს ან უცნობი აბრევიატურის მნიშვნელობას — და თითოეული თავდაჯერებით იგონებს განსხვავებულ პასუხს, არცერთი სწორი არაა. ინდუსტრიაში ამას *ჰალუცინაცია*-ს უწოდებენ, თითქოს აღქმის დარღვევა იყოს. ნაშრომის პირველი ნაბიჯი სწორედ ამ იდუმალების მოხსნაა.

დავიწყოთ იმით, როგორ იქმნება მოდელი. სწავლების პირველ და ყველაზე დიდ ეტაპზე ის უზარმაზარი ტექსტის კორპუსიდან სწავლობს, პრაქტიკულად, როგორ გამოიყურება გამართული ენა. ახლა ავიღოთ ფაქტი, რომელსაც რაიმე შიდა კანონზომიერება არ აქვს — ერთი კონკრეტული ადამიანის დაბადების თარიღი. თუ ეს თარიღი სასწავლო ტექსტში ერთხელ გამოჩნდა ან საერთოდ არ გამოჩნდა, pattern-learner-ს დასაჭერი არაფერი აქვს: მოდელის თვალით პასუხი თვითნებურია. ავტორები ამას ძველი იდეის — Alan Turing-ის სხვა პრობლემისთვის გამოყენებული მეთოდის — მეშვეობით ზუსტად აყალიბებენ: თუ დაბადების თარიღების მეხუთედი მონაცემებში მხოლოდ ერთხელ გვხვდება, მოდელისგან უნდა ველოდოთ, რომ სულ მცირე დაახლოებით მეხუთედზე შეცდება — არა იმიტომ, რომ „გაფუჭებულია“, არამედ იმიტომ, რომ სასწავლად საკმარისი ინფორმაცია თავიდანვე არ არსებობდა. (ამავე ლოგიკით, ქვეყნების დედაქალაქები თითქმის არასოდეს ერევა: ისინი ტექსტებში მუდმივად მეორდება.) ავტორები ამტკიცებენ, რომ ჭეშმარიტი წინადადების დამაჯერებელი სიცრუისგან გარჩევა თავისთავად რთული ამოცანაა, ხოლო *მხოლოდ* ჭეშმარიტი წინადადებების გენერირება სულ მცირე ამდენივე სირთულეს მოითხოვს. შეცდომის გარკვეული იატაკი სისტემაში თავიდანვეა.

> **ძირითადი იდეა: როგორ დაითვალო ის, რაც ჯერ არ გინახავს**
>
> ეს მართლაც ჭკვიანურ, ენობრივ მოდელებზე ბევრად ძველ იდეას ეყრდნობა. **წარმოიდგინეთ ფერადი ბურთების ტომარა.** არ იცით რამდენი ფერია შიგნით. ზედიზედ იღებთ 100 ბურთს და ითვლით: წითელი 40, ლურჯი 25, მწვანე 15, ყვითელი 5, იისფერი 3, ნარინჯისფერი 2 — და შემდეგ **ათი განსხვავებული ფერი, რომლებიც თითო-თითოჯერ შეგხვდათ.**
> Turing-ის რეალური კითხვა, სულ სხვა პრობლემაში, ასეთი იყო: *რა არის ალბათობა, რომ შემდეგი ბურთის ფერი ჯერ საერთოდ არ გინახავთ?* რაც არასოდეს ამოგიღიათ, პირდაპირ ვერ დათვლით — მაგრამ შეგიძლიათ დაითვალოთ ფერები, რომლებიც მხოლოდ ერთხელ ნახეთ, ე.წ. „singletons“. **Good–Turing estimation**-ის ხრიკი ისაა, რომ ყველა ამოღებულ ბურთში singleton-ების წილი დაახლოებით აფასებს იმ ალბათობას, რომელიც ჯერ უნახავ ფერებში „იმალება“. ასიდან ათი ამოღება ერთჯერადი ფერი იყო, ამიტომ

> შემდეგი ბურთის სრულიად ახალი ფერის ალბათობა დაახლოებით **10 / 100 = 10%**-ია.
> ერთხელ ნანახი ფერები *შეცდომები* არაა. ისინი ჩვენი უცოდინარობის საზომია: ბევრი singleton ნიშნავს, რომ სამყაროში კიდევ ბევრი ისეთი რამაა, რომელიც ნიმუშში უბრალოდ ჯერ არ შეგხვედრიათ.
> **ახლა ფერები დაბადების თარიღებით შეცვალეთ,** ტომარა კი მოდელის სასწავლო ტექსტით. ვთქვათ, ნანახ დაბადების თარიღებში **ყოველი მეხუთე მხოლოდ ერთხელ ჩანს**. იგივე ლოგიკა გვეუბნება, რომ დაახლოებით მეხუთედი ალბათობა ისეთ თარიღებზე მოდის, რომლებიც მოდელს პრაქტიკულად არასოდეს უნახავს — და დაბადების დღეს დასკვნით ვერ „გამოითვლი“. ერთხელ ნანახ ან სრულიად უნახავ თარიღზე მოდელს საიმედო საფუძველი არ აქვს და დაახლოებით **ყოველ მეხუთე** შემთხვევაში შეცდება. დამატებითი „ჭკუა“ ამას ვერ ასწორებს: სასწავლი ინფორმაცია არ არსებობდა.
> ეს მთელი არგუმენტის მინიატურაა: singleton სიჩქარე გვიჩვენებს, მოცემული მონაცემებიდან მსოფლიოს რა ნაწილი რჩება უსწავლელი, და ეს შეცდომების **ქვედა ზღვარს** ქმნის. ამიტომაც მოდელი დედაქალაქებს თითქმის არ ცდება — *Paris* ტექსტებში გამუდმებით გვხვდება, singleton სიჩქარე თითქმის ნულია და სასწავლი მასალა უამრავია.

ეს ხსნის, საიდან მოდის hallucination-ები. მაგრამ არ ხსნის, რატომ *რჩება* ისინი მოგვიანებითაც — რატომ bluff-ობს მოდელი მაშინაც, როცა შემდგომი სწავლება დახმარებისა და პატიოსნების გასაუმჯობესებლად ტარდება, ნაცვლად იმისა, რომ ეჭვი აღიაროს. აქ ნაშრომის ანალოგია უხერხულად ზუსტია. წარმოიდგინეთ სტუდენტი გამოცდაზე, რომელმაც პასუხი არ იცის. თუ ცარიელი პასუხი ნულ ქულას იღებს, გამოცნობილი პასუხი კი შეიძლება ერთ ქულად იქცეს, ქულის მაქსიმიზაციის სტრატეგია გამოცნობაა — კონკრეტულად და თავდაჯერებით, არა „არ ვიცი“. სტუდენტები ამას სწავლობენ. როგორც ჩანს, მოდელებიც — რადგან ჩვენც ასე ვაფასებთ. ავტორებმა შეისწავლეს ის ბენჩმარკები და leaderboard-ები, რომლებზეც სფერო რეალურად კონკურირებს, და ნახეს, რომ თითქმის ყველგან „არ ვიცი“ ზუსტად იმავე ქულას იღებს, რასაც არასწორი პასუხი: ნულს. ასეთი წესით მოდელი, რომელიც ყოველთვის გამოიცნობს, მოუგებს სხვაგვარად იდენტურ მოდელს, რომელიც საკუთარ გაურკვევლობას პატიოსნად აფიქსირებს. გარკვეული აზრით, ჩვენ მათ ამას ქულებით ვასწავლით.



ბენჩმარკმა გამოცნობა შეიძლება რაციონალურ სტრატეგიად აქციოს. გავრცელებული შეფასება-ის პირობებში (მარცხნივ) არასწორი პასუხიც და პატიოსანი „არ ვიცი“-ც ნულს იღებს — ამიტომ გამოცნობას მხოლოდ მოგება შეუძლია. თუ წესს თავად კითხვაში ჩაწერთ და შეცდომას ჯარიმას დაუწესებთ (მარჯვნივ), გაურკვევლობის დროს თავის შეკავება უკეთესი არჩევანი ხდება. ეს ცვლის იმას, რას აჯილდოებს ტესტი; თავისთავად hallucination-ს არ წყვეტს.

Original diagram — The Clean Paper CC BY 4.0

ესაა ნაწილი, რომელიც ყველაზე მეტად ღირს დასამახსოვრებლად, რადგან გავრცელებულ სათაურს ეწინააღმდეგება. ჰალუცინაცია ხშირად ტექნოლოგიის გარდაუვალ, თითქმის მისტიკურ ზღვარად იყიდება. ნაშრომი ორივე კომპონენტს ედავება. pretraining-ის შეცდომის იატაკი საიდუმლო არაა — ეს ჩვეულებრივი სტატისტიკური შეცდომაა, რომელსაც machine learning ათწლეულებია იცნობს. ხოლო მოგვიანებით მისი შენარჩუნება სრულად გარდაუვალი არ არის — ნაწილობრივ ეს ჩვენივე შექმნილი სტიმულია, რომლის შეცვლაც შეგვიძლია. სისტემა, რომელიც ეჭვის დროს უბრალოდ პასუხზე უარს იტყოდა, ცრუ ფაქტს აღარ მოიგონებდა; deployed მოდელები ასე იმიტომ არ იქცევიან, რომ ჩვენი scoreboards ხშირად უარს სჯის.

რა გააკეთეს ავტორებმა

ნაშრომი სამი ნაწილისგან შედგება. პირველი არის მათემატიკური არგუმენტი, რომ pretraining-ის დროს გარკვეული ჰალუცინაცია სტატისტიკურად გარდაუვალია: „მხოლოდ ვალიდური ტექსტის გენერირება“ სულ მცირე ისეთივე რთულია, როგორც ბინარული კლასიფიკაციის ამოცანა — „ეს განცხადება ვალიდურია თუ არა?“. მეორეა

არგუმენტი, რომელსაც ათი გავლენიანი ბენჩმარკის გამოკითხვა უჭერს მხარს, რომ გავრცელებული სიზუსტე-მეტრიკები გამოცნობას abstention-ზე აჯილდოებს. მესამეა შემოთავაზებული გამოსწორება და მისი case კვლევა: **ღია-rubric** შეფასებები, სადაც შეფასება თავად კითხვაშია გაწერილი (მაგალითად, „სწორი პასუხი +1, არასწორი −1; თუ 50%-ზე ნაკლებად დარწმუნებული ხარ, თავი შეიკავე“), რათა მოდელმა იცოდეს, როდის ჯილდოვდება პატიოსნება. ამას ოთხ frontier მოდელზე ამოწმებენ — Google Gemini 3 Pro, OpenAI GPT-5, xAI Grok 4 და Anthropic Claude Opus 4.5 — SimpleQA-ის 4,326 ფაქტობრივ კითხვაზე. ავტორები პირდაპირ ამბობენ, რომ ეს მხოლოდ საილუსტრაციო case კვლევაა და „არა კონტროლირებული შეფასება მოდელებს შორის“: default პარამეტრები, tuning-ის გარეშე და ხარჯის ნორმალიზაციის გარეშე.

რა აღმოაჩინეს

- **წინასწარი წვრთნა გარკვეულ შეცდომას აიძულებს.** მოდელის თავდაჯერებული სიცრუის სიხშირეს ქვედა ზღვარი აქვს, რომელიც დაახლოებით უკავშირდება მისგან აგებული საუკეთესო „ეს განცხადება ვალიდურია?“ კლასიფიკატორის შეცდომას. ფაქტებისთვის, რომლებსაც სასწავლი კანონზომიერება არ აქვთ, ეს იატაკი სულ მცირე *singleton სიჩქარე*-ია — იმ ფაქტების წილი, რომლებიც სასწავლო მონაცემებში ერთხელ ჩანს. სრულად სუფთა მონაცემებითაც გარკვეული ჰალუცინაცია გარდაუვალია.
- **შეფასება გამოცნობას კონკრეტულად აჯილდოებს.** ჩვეულებრივ correct/incorrect შეფასება-ში არასოდეს abstain-ება ოპტიმალური სტრატეგიაა, ხოლო ავტორების გამოკითხვაში პოპულარული ბენჩმარკების უმეტესობა „არ ვიცი“-ს უბრალოდ არასწორ პასუხად აფასებს. საკუთარი მოდელებიდან თვალსაჩინო მაგალითიც მოჰყავთ: SimpleQA-ზე raw სიზუსტე ოდნავ *უპირატესობას აძლევს* OpenAI o4-mini-ს, რომელიც თითქმის ყველაფერს პასუხობს და სამ მეოთხედზე მეტ შემთხვევაში ცდება, GPT-5-mini-სთან შედარებით, რომელიც ბევრად ნაკლებ შეცდომას უშვებს, რადგან ეჭვისას თავს იკავებს. scoreboard-ზე უფრო დაუფიქრებელი მოდელი უკეთ ჩანს.
- **ღია rubric-ები სტიმულს ატრიალებს — მათ case კვლევაში.** ავტორები ამოწმებენ მარტივ mitigation-ს: მოდელს ორჯერ უპასუხებინონ და თუ პასუხები არ ემთხვევა, abstain გააკეთოს. სტანდარტული სიზუსტე-ის პირობებში ეს შეცდომებს ამცირებს, მაგრამ სიზუსტე-საც ამცირებს — ანუ მეტრიკა მის გამოყენებას აფერხებს. ღია rubric-ებში იგივე mitigation ოთხივე მოდელზე სხვადასხვა penalty-ის პირობებში უკეთ გამოდის; GPT-5-mini-ც, რომელსაც raw სიზუსტე გაურკვევლობის აღიარებისთვის სჯიდა, o4-mini-ს უსწრებს, როცა შეფასება წესები წინასწარ და ღიადაა მოცემული (თითო მოდელზე n = 4,326 კითხვა).

რას ნიშნავს ეს, სავარაუდოდ

hallucination-ის შემცირება, ავტორების აზრით, ძირითადად კიდევ უფრო მეტი სპეციალური ჰალუცინაცია ტესტის გამოგონებას კი არ მოითხოვს, არამედ mainstream საკონტროლო ნიშნული-ების გაურკვევლობა შეფასება-ის შეცვლას, რათა „არ ვიცი“-ს თქმა აღარ ისჯებოდეს. სანამ scoreboard არ შეიცვლება, hallucination-ის შემცირება მოდელს სიზუსტე-ქულებს დააკლებს და მისი დანერგვა სისტემურად ნაკლებად მომგებიანი დარჩება. ამიტომ ავტორები პრობლემას „socio-technical“-ს უწოდებენ: საჭიროა როგორც უკეთესი მეტრიკა, ისე გავლენიანი leaderboard-ების მიერ მისი მიღება.

რას *არ* ამტკიცებს ეს

- ნაშრომი **არ აჩვენებს**, რომ ღია rubric-ები რეალურ deployed სისტემებში hallucination-ს აგვარებს. მხარდამჭერი ექსპერიმენტი მცირე და შეგნებულად **არაკონტროლირებული** case კვლევაა — ოთხი მოდელი default პარამეტრებით, ერთი mitigation, ერთი factual-QA ტესტი — რომლის მიზანია *სტიმულის შემობრუნების* დემონსტრირება და არა მოდელების რეიტინგი ან ზოგადი ეფექტიანობის დამტკიცება.
- ის **არ ამბობს**, რომ grading ერთადერთი მიზეზია. სასწავლო მონაცემების შეცდომები, რეალურად რთული ამოცანები და უცნობი prompts ცალკე წყაროებად რჩება.
- ის **არ უჭერს მხარს** პოპულარულ მოსაზრებას, რომ ჰალუცინაცია გარდაუვალია. ავტორები საპირისპიროს ამტკიცებენ: სისტემა, რომელიც მხოლოდ შესამოწმებელ კითხვებს უპასუხებდა და სხვა შემთხვევაში „არ ვიცი“-ს იტყოდა, ცრუ ფაქტობრივ პასუხებს არ გამოიგონებდა.
- ის **არ აქრობს** pretraining-ის შეცდომა floor-ს — ნაშრომი მას ხსნის და საზღვრავს; ეს ქვედა ზღვარი თავდაჯერებულ ფაქტობრივ შეცდომებს ეხება და არა მოდელის ყველა ტიპის ქცევას.
- ის **არ აჩვენებს**, რომ ღია rubric თავისთავად *საკმარისია*. იგი ცვლის შეფასების სტიმულს და retrieval-ის, ინსტრუმენტი use-ის ან უკეთ calibrated მოდელების შემცვლელი არ არის.

რამდენად ძლიერია მტკიცებულება?

- ბირთვი **მათემატიკაა** — ფორმალური ქვედა საზღვრები და არა ემპირიული გაზომვები. თეორიული არგუმენტი თავის დაშვებებში მყარად დგას.
- ის ეყრდნობა პრობლემის **განზრახ გამარტივებულ მოდელებს**; ავტორები თავად აღნიშნავენ „false trichotomy“-ს, როცა ყველა პასუხი მხოლოდ სწორად, არასწორად ან „არ ვიცი“-დ იყოფა, და იდეალიზებულ „arbitrary facts“ გარემოს, სადაც ყველაზე სუფთა საზღვარი მიიღება.
- საკონტროლო ნიშნული გამოკითხვა არის **პატარა, შერჩეული ნიმუში** — ათი

გავლენიანი შეფასება და არა სრული აუდიტი.
- case კვლევა **რეალურია, მაგრამ შეზღუდული**: ოთხი frontier მოდელი, ერთი mitigation, მხოლოდ SimpleQA, default settings და ავტორებისავე ფორმულირებით „არა კონტროლირებული შეფასება“. ეს incentive-არგუმენტის დამტკიცება of ცნებაია და არა საკონტროლო ნიშნული-რეზულტატი.
- მნიშვნელოვანია ხედვის წერტილიც: ოთხი ავტორიდან სამი **OpenAI-ში მუშაობს ან მუშაობდა**, ხოლო ნაშრომი ამტკიცებს, რომ სფერომ მოდელების შეფასების წესი უნდა შეცვალოს. ეს დაინტერესებული მხარის კარგად დასაბუთებული პოზიციაა და არა ნეიტრალური გარე მიმოხილვა — გასათვალისწინებელი, მაგრამ არა ავტომატურად უარსაყოფი. ნაშრომის სასარგებლოდ ისიც უნდა ითქვას, რომ საკუთარი მოდელების — o4-mini-სა და GPT-5-mini-ს — შეფასების პრობლემას ისევე აკრიტიკებს, როგორც სხვებისას.

რატომ არის მნიშვნელოვანი

ნაშრომი ზედმეტად ჰაიპირებულ პრობლემას უფრო ჩვეულებრივ ჩარჩოში აბრუნებს. „ჰალუცინაცია“ ხშირად ან მისტიკურ დეფექტად, ან შეუძრავ კედლადაა წარმოდგენილი; აქ ის ნაწილობრივ ჩვეულებრივი სტატისტიკური ზღვარია, რომლის გაგებაც შეგვიძლია, და ნაწილობრივ ჩვენივე არჩეული სტიმული. ფართო გაკვეთილი უფრო მშვიდი და სასარგებლოა: reliability-ის შემდგომი პროგრესი შეიძლება იმდენადვე იყოს დამოკიდებული *იმაზე, რას ვზომავთ*, რამდენადაც იმაზე, რას ვაშენებთ.

მოკლე შეჯამება

ენობრივი მოდელების თავდაჯერებული ცრუ პასუხები ორი წყაროდან მოდის. პირველი სტატისტიკურია: როცა ფაქტს სასწავლი კანონზომიერება არ აქვს, ენის მიბაძვაზე გაწვრთნილი მოდელი ზოგჯერ შეცდება და ამ შეცდომის ქვედა ზღვარი შეიძლება შეფასდეს — მაგალითად, იმ ფაქტების წილით, რომლებიც სასწავლო მონაცემებში მხოლოდ ერთხელ ჩანს. მეორე არის სტიმული: თითქმის ყველა მნიშვნელოვანი საკონტროლო ნიშნული „არ ვიცი“-ს იმავე ქულას აძლევს, რასაც არასწორ პასუხს, ამიტომ გამოცნობა ყოველთვის უფრო მომგებიანია — იმდენად, რომ მოდელი, რომელიც შემთხვევათა სამ მეოთხედზე მეტში ცდება, შეიძლება scoreboard-ზე უფრო პატიოსან, ხშირად abstain-ებულ მოდელს აჯობოს. ავტორების წინადადებაა არა კიდევ ერთი hallucination-ტესტი, არამედ „ღია rubrics“: შეფასება-ის წესი თავად კითხვაში ჩაწერეთ. ოთხ frontier მოდელზე ჩატარებულ მცირე case კვლევაში ეს სტიმულს აბრუნებს და hallucination-ის შემამცირებელ მეთოდს აჯილდოებს ნაცვლად დასჯისა. ეს theory-plus-გამოკითხვა ნაშრომია პატარა, პირდაპირ აღწერილი როგორც არაკონტროლირებული ექსპერიმენტით; peer-reviewed ვერსია *Nature*-ში გამოქვეყნდა. გამოსწორება პერსპექტიულია, მაგრამ მასშტაბურად ჯერ არ არის დადასტურებული, ხოლო ჰალუცინაცია არც მისტიკური და არც მკაცრად გარდაუვალი მოვლენაა.

ზედმეტი ხმაურის გარეშე

რას აჩვენებს ნაშრომი: მათემატიკურ ქვედა ზღვარს, რომლის პირობებშიც pretraining-ისას გარკვეული ჰალუცინაცია იძულებით ჩნდება (pattern-ის არმქონე ფაქტებისთვის სულ მცირე „singleton სიჩქარე“); გამოკითხვას, სადაც წამყვანი საკონტროლო ნიშნული-ების უმეტესობა „არ ვიცი“-ს ქულას არ აძლევს; და ოთხმოდელიან case კვლევას, სადაც შეფასება-ის prompt-ში ღიად ჩაწერა („ღია rubrics“) hallucination-ის შემამცირებელ მეთოდს იმარჯვებინებს იქ, სადაც ჩვეულებრივი სიზუსტე მას სჯიდა.

რა არის სარწმუნო, მაგრამ დაუმტკიცებელი: რომ mainstream საკონტროლო ნიშნული-ებში ღია rubric-ების ფართოდ დანერგვა deployed მოდელებში hallucination-ს მნიშვნელოვნად შეამცირებს. მხარდამჭერი ექსპერიმენტი მცირე და პირდაპირ აღწერილია როგორც არაკონტროლირებული.

რას არ აჩვენებს: რომ ჰალუცინაცია გარდაუვალია (საპირისპიროს ამტკიცებს); რომ grading ერთადერთი მიზეზია; რომ ჰალუცინაცია მთლიანად შეიძლება ავტომატურად გაქრეს; ან რომ case კვლევა ოთხ მოდელს ერთმანეთის წინააღმდეგ რეიტინგავს.

მთავარი შეზღუდვები: განზრახ გამარტივებული correct/incorrect/„არ ვიცი“ მოდელი (ავტორები თვითონ უწოდებენ მას „false trichotomy“-ს); ათი შეფასების მცირე შერჩეული გამოკითხვა; არაკონტროლირებული case კვლევა (ოთხი მოდელი, ერთი mitigation, ერთი ტესტი, default settings); და OpenAI-სთან დაკავშირებული ავტორების წინადადება იმის შესახებ, როგორ უნდა შეიცვალოს სფეროს შეფასების პრაქტიკა.

რამდენად უნდა ენდოს ამას ზოგადი მკითხველი? მაღალი ნდობა იმისა, რომ ჰალუცინაცია არც მისტიკურია და არც მკაცრად გარდაუვალი, და რომ mainstream საკონტროლო ნიშნული-ები დღეს ხშირად გამოცნობას აჯილდოებს. საშუალო ნდობა შემოთავაზებულ გამოსწორებაზე — დამტკიცება of ცნება რეალურია, მაგრამ ფართო და მასშტაბური ეფექტიანობა ჯერ არ არის ნაჩვენები.

წყაროები

Nature 653, 1047–1050 (2026)
https://doi.org/10.1038/s41586-026-10549-w

Why Language Models Hallucinate (arXiv 2509.04664)
https://arxiv.org/abs/2509.04664

hallucinations-paper-experiments (OpenAI)
https://github.com/openai/hallucinations-paper-experiments

SimpleQA
https://github.com/openai/simple-evals