The Clean Paper

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

Kodėl kalbos modeliai haliucinuoja — ir kodėl mūsų vertinimo būdas tai palaiko

Užtikrintai pateikiami klaidingi teiginiai, kuriuos vadiname „haliucinacijomis“, nėra paslaptinga sistemos klaida: dalis jų yra statistinis mokymo šalutinis produktas, o išlieka jie todėl, kad įprasti testai už drąsų spėjimą atlygina labiau nei už sąžiningą „nežinau“. Keturių pažangiausių modelių atvejo analizė rodo, kad vertinimo taisykles aiškiai įrašius į užklausą („atviros vertinimo taisyklės“), ši paskata apsiverčia.

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

Kodėl modeliai spėja — ir kodėl mes juos to išmokėme

Paklauskite didelio kalbos modelio, kada gimė nepažįstamas žmogus, ir jis gali ramiai, tarsi skaitydamas iš kortelės, atsakyti „kovo 7-ąją“ — ir suklysti tris kartus iš eilės, kiekvienąkart nurodydamas vis kitą datą. Autoriai pateikia būtent tokio pobūdžio pavyzdžių: pirmaujantys modeliai, gavę paprastą faktinį klausimą — žmogaus gimimo datą ar neaiškios santrumpos reikšmę — kiekvienas užtikrintai sugalvoja skirtingą atsakymą, ir nė vienas nėra teisingas. Pramonėje tai vadinama *haliucinacija*, tarsi būtų suvokimo sutrikimas. Pirmasis straipsnio žingsnis — parodyti, kad čia nėra nieko mistiško.

Pradėkime nuo to, kaip modelis sukuriamas. Pirmajame ir didžiausiame mokymo etape jis iš milžiniško tekstų kiekio iš esmės mokosi, kaip atrodo sklandi kalba. Dabar paimkime faktą, kuriame nėra jokio dėsningumo — konkretaus žmogaus gimimo datą. Jei ši data mokymo tekstuose pasirodė vieną kartą arba visai nepasirodė, dėsningumų besimokančiam modeliui nėra už ko užsikabinti: jo požiūriu atsakymas yra savavališkas. Autoriai tai suformuluoja tiksliau, pasitelkdami seną Alano Turingo idėją iš visai kitos problemos: jei kas penkta gimimo data duomenyse pasirodo tik kartą, turėtume tikėtis, kad modelis suklys bent dėl maždaug kas penktos tokios datos — ne todėl, kad modelis sugedęs, o todėl, kad mokytis tiesiog nebuvo iš ko. (Dėl tos pačios priežasties modeliai beveik neklysta dėl valstybių sostinių: šie faktai tekstuose kartojasi nuolat.) Autoriai atsargiai argumentuoja, kad atskirti teisingą teiginį nuo įtikinamai skambančio klaidingo teiginio jau savaime yra sunki užduotis, o generuoti *vien tik* teisingus teiginius yra bent ne lengviau. Taigi tam tikras klaidų minimumas yra neišvengiamas.

Pagrindinė idėja: kaip suskaičiuoti tai, ko dar nematėte

Čia remiamasi išties gudria idėja, gerokai senesne už kalbos modelius ir verta ją suprasti atskirai.

Pradėkite nuo maišelio su spalvotais rutuliukais. Nežinote, kiek jame yra spalvų. Ištraukiate 100 rutuliukų, po vieną, ir suskaičiuojate: 40 raudonų, 25 mėlynus, 15 žalių, 5 geltonus, 3 violetinius, 2 oranžinius — ir dar **dešimt skirtingų spalvų, kurių kiekviena pasirodo lygiai vieną kartą.**

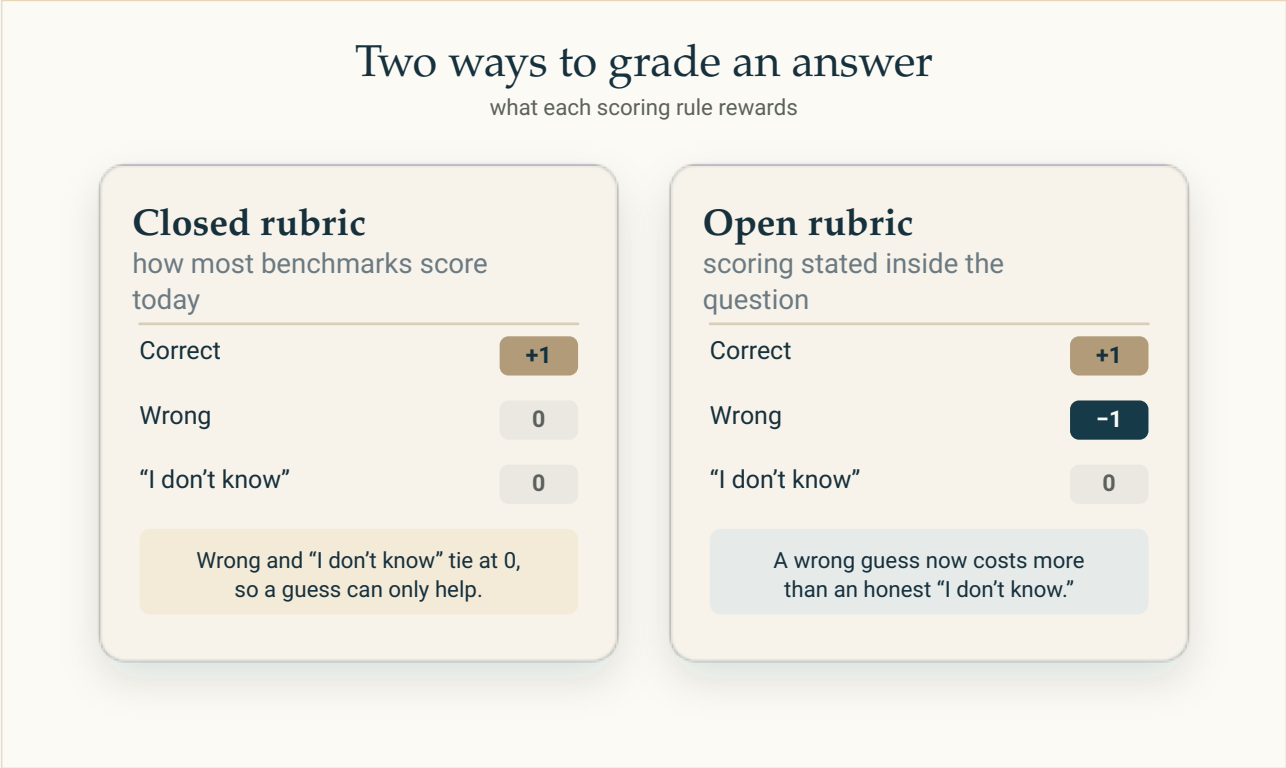
Turingui visai kitoje problemoje kilo toks klausimas: *kokia tikimybė, kad kitas rutuliukas bus tokios spalvos, kurios dar nė karto nematėte?* To, ko niekada neištraukėte, tiesiogiai suskaičiuoti negalite. Tačiau **galite** suskaičiuoti spalvas, pasirodžiusias tik vieną kartą — vadinamuosius vienetinius atvejus (*singletons*). **Goodo–Turingo įvertinimo** gudrybė tokia: vienetinių atvejų dalis imtyje leidžia įvertinti tikimybės masę, slypinčią dar nematytose spalvose. Jei 10 iš 100 ištrauktų rutuliukų buvo vienkartinių spalvų, tikimybė, kad kitas rutuliukas bus visiškai naujos spalvos, yra maždaug **10 / 100 = 10 %.**

Tos vieną kartą matytos spalvos nėra *klaidos*. Jos matuoja jūsų nežinojimą: kai imtyje daug spalvų pasirodo po vieną kartą, pati imtis signalizuoja, kad pasaulyje yra dar daugiau variantų, kurių paprasčiausiai nepaėmėte.

Dabar spalvas pakeiskite gimimo datomis, o maišelį — modelio mokymo tekstais. Tarkime, kad iš modelio matytų gimimo datų **kas penkta pasirodė lygiai vieną kartą.** Veikia ta pati logika: maždaug penktadalis tikimybės tenka datoms, kurių modelis praktiškai nėra matęs, o gimimo data neturi dėsningumo, iš kurio ją būtų galima išvesti. Vadinasi, vieną kartą matyta

> arba niekada nematyta data yra tarsi moneta, kurios modelis negali tinkamai pasverti, ir maždaug **kas penktu** tokiu atveju jis suklys. Jokios papildomos gudrybės to nepanaikina: mokytis tiesiog nebuvo iš ko.
> Tai ir yra visas argumentas sumažintas iki esmės: vienetinių atvejų dažnis parodo, kokia pasaulio dalis yra neišmokstama *iš šių duomenų*, ir tai tampa **apatine klaidų riba**. Dėl tos pačios priežasties modelis beveik niekada nesuklysta dėl sostinės — *Paryžius* tekstuose pasirodo nuolat, vienetinių atvejų dalis čia artima nuliui, todėl mokymuisi medžiagos gausu.

Tai paaiškina, iš kur haliucinacijos atsiranda. Tačiau nepaaiškina, kodėl jos *išlieka* — kodėl po visų vėlesnių mokymo etapų, skirtų modeliui padaryti naudingesniam ir sąžiningesniam, jis vis tiek blefuoja, užuot pripažinęs abejonę. Čia straipsnio analogija beveik nemaloniai tiksli. Įsivaizduokite studentą per egzaminą, nežinantį atsakymo. Jei tuščias laukelis vertas nulio, o spėjimas gali duoti vieną tašką, pažymį maksimalizuojanti strategija yra spėti — užtikrintai ir konkrečiai, niekada neatsakyti „nesu tikras“. Studentai tai išmoksta. Pasirodo, modeliai irgi — nes mes juos vertiname taip pat. Autoriai peržiūrėjo testus, pagal kuriuos sritis realiai konkuruoja ir kurių reitinguose modeliai bando kilti, ir nustatė, kad beveik visi už „nežinau“ skiria lygiai tiek pat, kiek už klaidingą atsakymą: nulį. Pagal tokią taisyklę visada spėjantis modelis aplenks kitą, visais kitais atžvilgiais tokį pat modelį, kuris sąžiningai pažymi savo neapibrėžtumą. Gana tiesiogine prasme mūsų vertinimo sistema moko juos spėlioti.



Vertinimo testai gali paversti spėjimą racionalia strategija. Pagal daugumoje testų taikomą sistemą (kairėje) ir klaidingas atsakymas, ir sąžiningas „nežinau“ vertinami nuliu, todėl spėjimas gali tik padėti. Jei pačiame klausime aiškiai nurodomos taisyklės ir už klaidą skiriama bauda (dešinėje), abejojant gali būti naudingiau susilaikyti nuo atsakymo. Tai pakeičia, už ką testas apdovanoja; savaime haliucinacijų problemos tai neišsprendžia.

Original diagram — The Clean Paper CC BY 4.0

Būtent šią dalį verta įsidėmėti, nes ji prieštarauja įprastai antraštei. Haliucinacijos dažnai pristatomos kaip neišvengiama, beveik mistinė technologijos riba. Straipsnis ginčija abi šias prielaidas. Pirminio mokymo klaidų minimumas nėra paslaptis — tai įprasta statistinė paklaida, kokią mašininis mokymasis nagrinėja jau dešimtmečius. Ir haliucinacijų išlikimas nėra neišvengiamas — iš dalies tai mūsų pačių sukurta paskata, kurią galėtume pakeisti. Sistema, kuri abejojanti tiesiog atsisakytų atsakyti, nekurstytų faktinių haliucinacijų; viena priežasčių, kodėl diegiami modeliai taip nesielgia, yra ta, kad mūsų reitingai už atsisakymą atsakyti baudžia.

Ką padarė autoriai

Straipsnį sudaro trys dalys. Pirmiausia pateikiamas matematinis argumentas, kad dalis haliucinacijų statistiškai neišvengiama jau pirminio mokymo metu: parodoma, jog užduotis „generuoti tik galiojantį tekstą“ yra bent tokia pat sunki kaip dvejetainė klasifikavimo užduotis „ar šis teiginys galiojantis?“. Antra, remiantis dešimties įtakingų testų apžvalga, argumentuojama, kad įprasti tikslumu grindžiami rodikliai apdovanoja spėjimą, o ne susilaikymą nuo atsakymo. Trečia, autoriai pasiūlo pataisą ir ją patikrina atvejo analizėje: **atviras vertinimo taisykles** (*open rubrics*), kai taškų skyrimo schema įrašoma tiesiai į klausimą, pavyzdžiui: „teisingas atsakymas vertas 1, klaidingas −1, todėl susilaikykite, jei esate mažiau nei 50 % tikri“. Taip modelis gali suprasti, kada už sąžiningą neapibrėžtumo pripažinimą bus atlyginta. Tai išbandoma su keturiais pažangiausiais modeliais — „Google“ Gemini 3 Pro, „OpenAI“ GPT-5, „xAI“ Grok 4 ir „Anthropic“ Claude Opus 4.5 — naudojant SimpleQA 4 326 faktinius klausimus. Autoriai aiškiai pabrėžia, kad tai iliustracinė atvejo analizė, o „ne kontroliuojamas modelių palyginimas“: naudotos numatytosios nuostatos, nebuvo derinimo ir sąnaudų normalizavimo.

Ką jie nustatė

- **Pirminis mokymas neišvengiamai palieka dalį klaidų.** Modelio užtikrintai skleidžiamų klaidingų teiginių dažnis turi apatinę ribą, susietą su geriausio iš to modelio sukonstruoto klasifikatoriaus „ar teiginys galiojantis?“ klaidų dažniu (apytikriai — maždaug dvigubu). Faktams, kuriuose nėra išmokstamo dėsningumo, ši riba yra bent *vienetinių atvejų dažnis* — mokymo duomenyse tik kartą pasirodančių faktų dalis. Todėl dalis haliucinacijų neišvengiama net ir visiškai švariuose duomenyse.
- **Vertinimo sistema konkrečiai apdovanoja spėjimą.** Taikant įprastą teisinga/klaidinga vertinimą, niekada nesusilaikyti nuo atsakymo yra optimali strategija, o autorių apžvalga rodo, kad didžioji dalis populiarių testų „nežinau“ tiesiog prilygina klaidai. Iškalbingas jų pačių pavyzdys: SimpleQA teste neapdorotas tikslumas šiek tiek *palankesnis* „OpenAI“ o4-mini, kuris atsako beveik į viską ir suklysta daugiau nei trimis ketvirtadaliais atvejų, negu GPT-5-mini, kuris klysta gerokai rečiau, nes abejojantis susilaiko. Reitinge beatodairiškesnis modelis atrodo geresnis.
- **Atviros taisyklės pakeičia paskatą (šioje atvejo analizėje).** Autoriai išbando paprastą hali-

ucinacijų mažinimo būdą: modelis atsako du kartus, o jei atsakymai nesutampa — susilaiko. Vertinant vien pagal įprastą tikslumą, šis metodas sumažina klaidų skaičių, *bet kartu sumažina ir tikslumo rodiklį,* todėl metrika jo diegimą atgraso. Taikant atviras taisykles tas pats metodas, esant įvairaus dydžio baudoms, pasirodo geresnis visiems keturiems modeliams; be to, GPT-5-mini, kurį neapdorotas tikslumas baudė už susilaikymą abejojant, aplenkia o4-mini, kai vertinimo taisyklės aiškiai pateiktos (n = 4 326 klausimai kiekvienam modeliui).

Ką tai greičiausiai reiškia

Haliucinacijų mažinimas daugiausia nėra dar vienų specialiai haliucinacijoms skirtų testų kūrimo klausimas. Svarbiau pakeisti, kaip pagrindiniai testai vertina neapibrėžtumą, kad prisipažinimas „nežinau“ nebebūtų baudžiamas. Kol reitingai nesikeis, haliucinacijų mažinimas ir toliau kainuos tikslumo taškus ir todėl bus atgrasomas. Dėl to autoriai problemą vadina „sociotechnine“: reikia ir geresnės metrikos, ir to, kad įtakingi reitingai ją iš tikrųjų pradėtų taikyti.

Ko tai *neįrodo*

- Tai **neparodo**, kad atviros taisyklės išsprendžia haliucinacijas realiomis naudojimo sąlygomis. Eksperimentinė atrama yra maža ir sąmoningai **nekontroliuojama** atvejo analizė — keturi modeliai su numatytosiomis nuostatomis, vienas pasirinktas mažinimo būdas ir vienas faktinių klausimų testas. Ji skirta parodyti *paskatos apsivertimą,* o ne surikiuoti modelius ar įrodyti bendrą veiksmingumą.
- Straipsnyje **neteigiama**, kad vertinimas yra *vienintelė* priežastis. Klaidos mokymo duomenyse, iš tiesų sunkios užduotys ir neįprastos užklausos išlieka atskirais klaidų šaltiniais.
- Straipsnis **nepalaiko** populiarios minties, kad haliucinacijos yra neišvengiamos. Autoriai teigia priešingai: sistema, atsakanti tik į patikrinamus klausimus, o kitais atvejais sakanti „nežinau“, niekada nepateiktų tokio tipo faktinės haliucinacijos.
- Tai **nepanaikina** pirminio mokymo klaidų ribos — tik ją paaiškina ir apriboja; ši riba taikoma užtikrintoms faktinėms klaidoms, o ne visam modelio elgesiui.
- Tai **neparodo**, kad vien atvirų taisyklių pakanka. Jos pakeičia, už ką vertinimo sistema skiria taškus, bet nepakeičia informacijos paieškos, įrankių naudojimo ar geresnio modelių kalibravimo.

Kiek tvirti yra įrodymai?

- Pagrindą sudaro **matematika** — formalios apatinės ribos, o ne empiriniai matavimai. Kaip teorinis argumentas, jis yra nuoseklus pagal savo prielaidas.
- Argumentas remiasi **sąmoningai supaprastintais problemos modeliais**. Patys autoriai pažymi „klaidingą trichotomiją“, kai kiekvienas atsakymas laikomas teisingu, klaidingu arba „nežinau“, ir

idealizuotą „savavališkų faktų“ aplinką, naudojamą aiškiausiai ribai išvesti.
- Testų apžvalga yra **nedidelė, atrinkta imtis** — dešimt įtakingų vertinimų, o ne išsamus auditas.
- Atvejo analizė yra **reali, bet ribota**: keturi pažangiausi modeliai, vienas mažinimo metodas, tik SimpleQA, numatytosios nuostatos ir aiškiai pasakyta, kad tai „ne kontroliuojamas vertinimas“. Tai paskatos argumento principo demonstracija, o ne naujas modelių reitingas.
- Verta įvardyti ir autorių poziciją: trys iš keturių autorių dirba arba dirbo **OpenAI**, o straipsnyje siūloma keisti visos srities modelių vertinimą. Tai gerai argumentuota suinteresuotos šalies pozicija, o ne neutralus išorinis vertinimas — ją reikia pasverti, o ne atmesti. Straipsnio naudai verta pažymėti, kad kritika lygiai taip pat taikoma ir pačios „OpenAI“ modeliams o4-mini bei GPT-5-mini.

Kodėl tai svarbu

Straipsnis kitaip suformuluoja pernelyg išpūstą problemą. „Haliucinacija“ dažnai pateikiama arba kaip paslaptingas defektas, arba kaip neįveikiama siena; čia ji tampa daug žemiškesnė ir iš dalies mūsų pačių sukurta — suprantama statistinė klaidų riba, ant kurios uždedame savo pasirinktą paskatų sistemą. Platesnė pamoka tylesnė, bet naudingesnė: tolesnė pažanga patikimumo srityje gali priklausyti ne mažiau nuo *to, ką matuojame*, negu nuo to, ką kuriame.

Trumpa santrauka

Užtikrintai pateikiami klaidingi kalbos modelių atsakymai kyla iš dviejų šaltinių. Pirmasis statistinis: kai faktas neturi išmokstamo dėsningumo, kalbą mėgdžioti mokytas modelis kartais suklys, o šią apatinę ribą galima įvertinti, pavyzdžiui, pagal tai, kiek faktų mokymo duomenyse pasirodo tik vieną kartą. Antrasis šaltinis — paskatos: beveik visi testai, pagal kuriuos modeliai reitinguojami, „nežinau“ vertina taip pat kaip klaidingą atsakymą, todėl spėti visada apsimoka — tiek, kad modelis, klystantis trimis ketvirtadaliais atvejų, gali aplenkti sąžiningesnį modelį, kuris abejojantis susilaiko. Autoriai siūlo ne dar vieną haliucinacijų testą, o „atviras vertinimo taisykles“: taškų skyrimą aiškiai įrašyti į patį klausimą. Keturių pažangiausių modelių atvejo analizėje tai pakeitė paskatą taip, kad haliucinacijas mažinantis metodas buvo apdovanotas, o ne nubaustas. Tai teorijos ir testų apžvalgos straipsnis su mažu, aiškiai nekontroliuojamu eksperimentu, recenzuotas *Nature*; siūloma pataisa perspektyvi, tačiau dar neparodyta, kad ji plačiai veikia dideliu mastu, o haliucinacijos, pasak autorių, nėra nei paslaptingos, nei absoliučiai neišvengiamos.

Be pagražinimų

Ką straipsnis parodo: matematinę apatinę ribą, dėl kurios dalis haliucinacijų pirminio mokymo metu neišvengiama (faktams be dėsningumo — bent „vienetinių atvejų dažnį“); apžvalgą, rodančią, kad dauguma pirmaujančių testų už „nežinau“ taškų neskiria; ir keturių modelių atvejo analizę, ku-

rioje į užklausą įrašytos vertinimo taisyklės („atviros taisyklės“) leido laimėti haliucinacijas mažinančiam metodui, kurį paprastas tikslumo rodiklis buvo baudęs.

Kas tikėtina, bet neįrodyta: kad į pagrindinius testus įtrauktos atviros taisyklės reikšmingai sumažintų haliucinacijas realiai naudojamuose modeliuose. Eksperimentinis pagrindas mažas ir aiškiai nekontroliuojamas.

Ko tai nerodo: kad haliucinacijos neišvengiamos (straipsnis teigia priešingai); kad vertinimo sistema yra vienintelė priežastis; kad haliucinacijas galima visiškai panaikinti; kad atvejo analizė leidžia tarpusavyje surikiuoti keturis modelius.

Svarbiausi apribojimai: sąmoningai supaprastintas modelis „teisinga / klaidinga / nežinau“ (autoriai jį vadina „klaidinga trichotomija“); nedidelė atrinkta testų apžvalga (dešimt vertinimų); nekontroliuojama atvejo analizė (keturi modeliai, vienas mažinimo būdas, vienas testas, numatytosios nuostatos); ir „OpenAI“ vadovaujamas argumentas apie tai, kaip visa sritis turėtų vertinti modelius.

Kiek tuo turėtų pasitikėti bendras skaitytojas? Labai tvirtai galima teigti, kad haliucinacijos nėra nei paslaptingos, nei absoliučiai neišvengiamos ir kad dabartiniai pagrindiniai testai skatina spėjimą. Vidutinis pasitikėjimas siūlomu sprendimu: jau yra reali principo demonstracija, bet dar nėra įrodymo, kad metodas plačiai ir dideliu mastu veikia.

Šaltiniai

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