The Clean Paper

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

Kāpēc valodas modeļi halucinē — un kāpēc mūsu vērtēšanas veids to uztur

Pārliecinošās nepatiesības, ko saucam par "halucinācijām", nav mistiska kļūme: daļa ir statistisks apmācības blakusefekts, bet tās saglabājas arī tāpēc, ka galvenie etaloni atalgo pārliecinošu minējumu vairāk nekā godīgu "nezinu". Gadījuma pētījumā ar četriem vadošiem modeļiem punktu noteikumu ierakstīšana uzvednē ("atklātas vērtēšanas shēmas") šo motivāciju apgriež.

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

Kāpēc modeļi min — un kāpēc mēs tos tam iemācījām

Pajautājiet lielam valodas modelim sveša cilvēka dzimšanas dienu, un tas var ar mierīgu pārliecību atbildēt "7. marts", it kā nolasītu datumu no kartītes — un trīs mēģinājumos trīsreiz kļūdīties ar trim dažādiem datumiem. Autori min tieši šādus piemērus: vadošajiem modeļiem uzdod vienkāršu faktu jautājumu — par cilvēka dzimšanas dienu vai reti lietota saīsinājuma nozīmi —, un katrs pārliecinoši izdomā citu atbildi, neviena nav pareiza. Nozarē to sauc par *halucināciju*, it kā tā būtu uztveres kļūme. Pirmais darba solis ir noņemt no šā procesa mistiku.

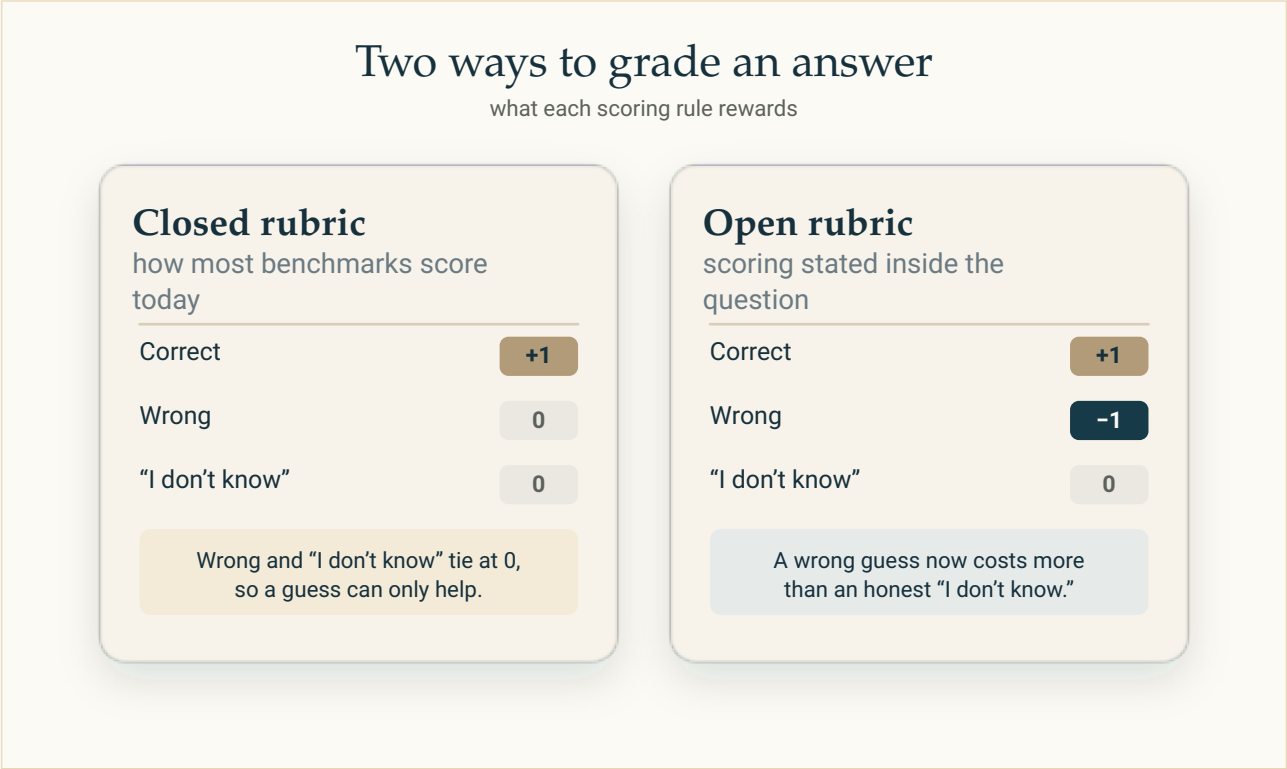
Sākam ar to, kā modelis tiek veidots. Pirmajā un lielākajā apmācības posmā tas būtībā no milzīga teksta daudzuma mācās, kā izskatās plūstoša valoda. Tagad paņemam faktu, kam nav aiz tā slēpta likumsakarība — viena konkrēta cilvēka dzimšanas datumu. Ja šis datums treniņtekstā parādījās vienreiz vai vispār neparādījās, rakstu apguvējam nav pie kā pieķerties: no modeļa skatpunkta atbilde ir patvaļīga. Autori šo intuīciju padara precīzu, aizņemoties senu ideju no Alana Tjūringa cita uzdevuma: ja viena piektdaļa dzimšanas datumu datos sastopama tikai vienreiz, modelim vajadzētu kļūdīties vismaz aptuveni vienā piektdaļā no tiem — ne tāpēc, ka modelis ir "salūzis", bet tāpēc, ka nebija ko iemācīties. (Tā paša iemesla dēļ modeļi gandrīz nekad nesajauc valstu galvaspilsētas: tās datos atkārtojas neskaitāmas reizes.) Autori rūpīgi argumentē, ka patiesa apgalvojuma atšķiršana no ticama nepatiesa apgalvojuma pati ir grūts uzdevums un ka tikai patiesu apgalvojumu ģenerēšana ir vismaz tikpat grūta. Kļūdu minimums ir iebūvēts pašā statistiskajā uzdevumā.

Pamatideja: kā saskaitīt to, ko vēl neesi redzējis

Šeit izmantota patiešām eleganta ideja, kas ir daudz vecāka par valodas modeļiem.
Iedomājieties maisu ar krāsainām lodītēm. Jūs nezināt, cik krāsu tajā ir. Izvelkat 100 lodītes un saskaitāt: sarkanas 40, zilas 25, zaļas 15, dzeltenas 5, violetas 3, oranžas 2 — un tad **desmit dažādas krāsas, no kurām katra parādās tieši vienreiz**.
Tjūringa jautājums citā problēmā bija šāds: *kāda ir varbūtība, ka nākamā lodīte būs krāsā, ko vēl nekad neesi redzējis?* To, ko nekad neesi izvilcis, tieši saskaitīt nevar. Taču var saskaitīt krāsas, kas parādījušās tieši vienreiz — "vienreizējos" gadījumus. **Good–Turing novērtējuma** triks ir tāds, ka šo vienreizējo novērojumu īpatsvars aptuveni novērtē varbūtības masu, kas slēpjas vēl neredzētajās kategorijās. Desmit no simt izvilktajām lodītēm bija vienreizējas krāsas, tātad iespēja, ka nākamā būs pavisam jauna krāsa, ir aptuveni **10 / 100 = 10%.**
Šīs vienreiz redzētās krāsas nav *kļūdas*. Tās mēra mūsu pašu nezināšanu: ja paraugā daudz kategoriju parādās tikai vienreiz, tas norāda, ka pasaulē ir vēl citas, kuras vienkārši neesam izvilkuši.
Tagad krāsas aizstājam ar dzimšanas dienām, bet maisu — ar modeļa treniņtekstu. Pieņemsim, ka no visiem redzētajiem dzimšanas datumiem **viens no pieciem parādās tikai vienreiz**. Tā pati metode saka: apmēram piektā daļa varbūtības masas atrodas datos, ko modelis faktiski nav pietiekami redzējis — un dzimšanas datumam nav vispārīgas likumsakarības, pēc kuras to varētu "izrēķināt". Tātad vienreiz vai nekad neredzēts datums ir fakts, kura pareizo vērtību modelis no raksta nevar atjaunot, un aptuveni **vienā no pieciem** gadījumiem tas kļūdīsies. Ar "gudrāku domāšanu" to nevar salabot: informācijas vienkārši nebija datos.
Tas ir visa argumenta kodols: vienreizējo gadījumu īpatsvars mēra, cik liela pasaules daļa

> *no konkrētajiem datiem* nav apgūstama, un tas veido **apakšējo robežu** kļūdām. Tāpēc modelis gandrīz nekad nekļūdās par galvaspilsētu: "Parīze" datos sastopama daudzkārt, tās vienreizējo novērojumu īpatsvars ir tuvu nullei, un modelim ir daudz, no kā mācīties.

Tas izskaidro, no kurienes halucinācijas rodas. Taču neizskaidro, kāpēc tās *saglabājas* pēc vēlākas apmācības, kuras mērķis ir padarīt modeli noderīgu un godīgu — kāpēc modelis joprojām blefo, nevis atzīst šaubas. Šeit autoru analoģija ir nepatīkami trāpīga. Iedomājieties studentu eksāmenā, kurš nezina atbildi. Ja tukša atbilde dod nulli, bet minējums var dot vienu punktu, atzīmes maksimizējošā stratēģija ir minēt — konkrēti un pārliecinoši, nevis teikt "neesmu drošs". Studenti to apgūst. Izrādās, modeļi arī, jo mēs tos vērtējam pēc līdzīgiem noteikumiem. Autori pārskatīja etalonus un līderu tabulas, pēc kurām nozare sacenšas, un secināja, ka gandrīz visi "nezinu" vērtē tieši tāpat kā nepareizu atbildi: ar nulli. Pie šāda noteikuma modelis, kas vienmēr min, būs labāks par citādi identisku modeli, kas godīgi norāda nenoteiktību. Burtiskā nozīmē mēs ar vērtēšanas sistēmu mudinām modeļus blefot.



Etaloni var padarīt minēšanu racionālu. Pie tipiskās vērtēšanas (pa kreisi) gan nepareiza atbilde, gan godīgs "nezinu" saņem nulli, tātad minējums var tikai palīdzēt. Ja noteikumus ieraksta pašā jautājumā un nepareizai atbildei paredz sodu (pa labi), nenoteiktības gadījumā izdevīgāk var kļūt atturēties no atbildes. Tas maina, ko tests atalgo; pats par sevi tas halucinācijas neatrisina.

Original diagram — The Clean Paper CC BY 4.0

Šī ir daļa, ko vērts paturēt prātā, jo tā pretojas parastajam virsrakstam. Halucinācijas bieži tiek pasniegtas kā neizbēgams, gandrīz mistisks tehnoloģijas ierobežojums. Darbs apstrīd abus vārdus. Priekšapmācības kļūdu minimums nav noslēpums — tā ir parasta statistiska kļūda, ko mašīnmācīšanās pazīst gadu desmitiem. Savukārt noturība nav pilnīgi neizbēgama — daļēji tā

ir mūsu pašu izveidota motivācija, ko var mainīt. Sistēma, kas vienkārši neatbildētu, kad nav pārliecināta, vispār nehalucinētu; iemesls, kāpēc izvietotie modeļi tā nerīkojas konsekventi, ir arī tas, ka mūsu rezultātu tabulas par atteikšanos soda.

Ko autori darīja

Darbam ir trīs daļas. Pirmkārt, matemātisks arguments, ka daļa halucināciju statistiski ir neizbēgama priekšapmācības laikā: uzdevums "ģenerēt tikai derīgu tekstu" tiek sasaistīts ar bināru klasifikācijas uzdevumu "vai šis apgalvojums ir derīgs?". Otrkārt, arguments, ko atbalsta desmit ietekmīgu etalonu pārskats, ka tipiskie precizitātes rādītāji atalgo minēšanu vairāk nekā atturēšanos no atbildes. Treškārt, piedāvāts labojums un gadījuma pētījums: **atklātas vērtēšanas shēmas** (*open rubrics*), kur punktu piešķiršanas noteikumi ir ierakstīti pašā jautājumā (piemēram, "pareiza atbilde +1, nepareiza −1; ja pārliecība zem 50%, atturies"), lai modelis zinātu, ka godīga nenoteiktība tiek atalgota. To pārbauda uz četriem vadošiem modeļiem — Google Gemini 3 Pro, OpenAI GPT-5, xAI Grok 4 un Anthropic Claude Opus 4.5 —, izmantojot SimpleQA 4326 faktu jautājumus. Autori skaidri uzsver, ka tas ir ilustratīvs gadījuma pētījums, "nevis kontrolēts modeļu salīdzinājums": izmantoti noklusētie iestatījumi, nav pielāgošanas un izmaksu normalizācijas.

Ko viņi atklāja

- **Priekšapmācība piespiež pieļaut daļu kļūdu.** Pārliecinošu nepatiesu apgalvojumu ģenerēšanas biežumam ir apakšējā robeža, kas saistīta ar labākā no modeļa iegūstamā "vai apgalvojums ir derīgs?" klasifikatora kļūdu. Fakta gadījumos bez iemācāmas likumsakarības šī robeža ir vismaz *vienreizējo gadījumu īpatsvars* — faktu daļa, kas treniņdatos parādās tieši vienreiz. Tātad pat perfekti tīros datos daļa kļūdu ir neizbēgama.
- **Vērtēšana konkrēti atalgo minēšanu.** Parastā pareizi/nepareizi shēmā nekad neatturēties ir optimāla stratēģija, un autoru aptaujā lielākā daļa populāru etalonu "nezinu" vienkārši ieskaita kā nepareizu atbildi. Spilgts piemērs no pašu OpenAI modeļiem: SimpleQA neapstrādātā precizitāte nedaudz *dod priekšroku* o4-mini, kas atbild gandrīz uz visu un kļūdās vairāk nekā trīs ceturtdaļās gadījumu, pār GPT-5-mini, kas kļūdās daudz retāk, jo nenoteiktības gadījumā biežāk atturas. Riskantākais modelis izskatās labāks līderu tabulā.
- **Atklāta vērtēšana maina motivāciju — vismaz gadījuma pētījumā.** Autori pārbauda vienkāršu halucināciju mazināšanas metodi: modelim jāatbild divreiz un jāatturas, ja abas atbildes nesakrīt. Pie parastās precizitātes kļūdu kļūst mazāk, bet arī precizitātes rezultāts krītas — tātad metrika attur no šā risinājuma. Pie atklātas vērtēšanas tas pats paņēmiens visiem četriem modeļiem kļūst izdevīgāks dažādos soda līmeņos; arī GPT-5-mini, ko neapstrādātā precizitāte sodīja par atturēšanos, apsteidz o4-mini, tiklīdz punktu noteikumi ir pateikti atklāti (n = 4326 jautājumi katram modelim).

Ko tas, visticamāk, nozīmē

Halucināciju samazināšana lielā mērā nav jautājums par arvien jaunu īpašu "halucināciju testu" izgudrošanu. Tas ir jautājums par to, kā galvenie etaloni novērtē nenoteiktību, lai "nezinu" vairs netiktu sodīts tāpat kā kļūda. Kamēr līderu tabula nemainās, halucināciju mazināšana bieži maksās precizitātes punktus un līdz ar to tiks netieši atturēta. Tāpēc autori problēmu sauc par "sociotehnisku": vajadzīgas gan labākas metrikas, gan ietekmīgo etalonu gatavība tās pieņemt.

Ko tas nepierāda

- Tas **neparāda**, ka atklātas vērtēšanas shēmas novērš halucinācijas reālos produktos. Eksperiments ir neliels, apzināti **nekontrolēts** gadījuma pētījums — četri modeļi ar noklusētiem iestatījumiem, viena mazināšanas metode un viens faktu jautājumu tests —, kura mērķis ir demonstrēt *motivācijas maiņu*, nevis sarindot modeļus vai pierādīt vispārīgu efektivitāti.
- Tas **neapgalvo**, ka vērtēšana ir *vienīgais* cēlonis. Treniņdatu kļūdas, patiesi grūti uzdevumi un nepazīstami jautājumi ir atsevišķi avoti.
- Tas **neatbalsta** populāro domu, ka halucinācijas ir neizbēgamas. Autori argumentē pretējo: sistēma, kas atbildētu tikai uz pārbaudāmiem jautājumiem un citādi teiktu "nezinu", nehalucinētu.
- Tas **neatceļ** priekšapmācības statistisko kļūdu minimumu — darbs to izskaidro un ierobežo; turklāt robeža attiecas uz pārliecinošām faktu kļūdām, ne visu modeļa uzvedību.
- Tas **neparāda**, ka atklātas vērtēšanas shēmas vienas pašas ir *pietiekamas*. Tās maina, ko novērtējums atalgo, bet neaizstāj informācijas izgūšanu, rīku izmantošanu vai labāku modeļa pārliecības kalibrēšanu.

Cik pārliecinoši ir pierādījumi?

- Kodols ir **matemātika** — formālas apakšējās robežas, nevis empīriski mērījumi. Teorētiskais arguments ir derīgs savu pieņēmumu robežās.
- Tas balstās **apzināti vienkāršotos problēmas modeļos**. Autori paši norāda uz "viltus trihotomiju", kur katra atbilde tiek klasificēta kā pareiza, nepareiza vai "nezinu", un izmanto idealizētu "patvaļīgu faktu" vidi, lai iegūtu tīrāko robežu.
- Etalonu apskats ir **neliels, atlasīts paraugs** — desmit ietekmīgi novērtējumi, nevis pilnīgs audits.
- Gadījuma pētījums ir **reāls, bet ierobežots**: četri vadoši modeļi, viena mazināšanas metode, tikai SimpleQA, noklusētie iestatījumi un autori to skaidri sauc par "ne kontrolētu novērtējumu". Tas ir motivācijas argumenta principa pierādījums, nevis modeļu etalons.
- Vērts nosaukt skatpunkta konfliktu: trīs no četriem autoriem strādā vai ir strādājuši **OpenAI**, bet raksts argumentē, ka nozarei jāmaina modeļu vērtēšana. Tas ir labi argumentēts ieinteresētas puses viedoklis, ne neitrāls ārējs audits — jāizvērtē, nevis jāignorē. Par labu darbam runā tas, ka kritika tikpat tieši attiecas uz pašu o4-mini un GPT-5-mini.

Kāpēc tas ir svarīgi

Darbs pārformulē ļoti pārspīlētu problēmu. “Halucinācija” parasti tiek pārdota vai nu kā mistisks defekts, vai kā nepārvarama siena; šeit tā kļūst par parastu un daļēji pašu radītu parādību — saprotamu statistisku kļūdu minimumu, kam virsū uzlikta mūsu pašu izvēlēta motivācija. Plašāka un noderīgāka mācība ir klusāka: turpmākais uzticamības progress var būt atkarīgs ne tikai no tā, *ko mēs būvējam*, bet arī no tā, *ko mēs mēram*.

Īss kopsavilkums

Pārliecinoši nepareizām valodas modeļu atbildēm ir divi galvenie avoti. Pirmais ir statistisks: ja faktam nav iemācāmas likumsakarības, valodas imitēšanai trenēts modelis dažkārt kļūdīsies, un šim minimumam var novērtēt apakšējo robežu, piemēram, pēc tā, cik faktu treniņdatos parādās tikai vienreiz. Otrais ir motivācija: gandrīz visi etaloni, pēc kuriem modeļus sarindo, “nezinu” vērtē tāpat kā nepareizu atbildi, tāpēc minēšana vienmēr var dot lielāku punktu skaitu — līdz pat situācijai, kur modelis, kas kļūdās trīs ceturtdaļās gadījumu, pārspēj godīgāku modeli, kas nenoteiktības gadījumā atturas. Autoru priekšlikums nav vēl viens halucināciju tests, bet “atklātas vērtēšanas shēmas”: punktu noteikumus ierakstīt pašā jautājumā. Gadījuma pētījumā ar četriem vadošiem modeļiem tas maina motivāciju tā, ka halucinācijas mazinoša metode tiek atalgota, nevis sodīta. Darbs apvieno teoriju, etalonu apskatu un nelielu, skaidri nekontrolētu eksperimentu; tas ir recenzēts *Nature*. Labojums ir daudzsološs, bet vēl nav pierādīts lielā mērogā, un halucinācijas darbā tiek raksturotas kā ne mistiskas, ne pilnīgi neizbēgamas.

Bez pārspīlējumiem

Ko pētījums parāda: matemātisku apakšējo robežu, pie kuras daļa halucināciju priekšapmācībā ir statistiski neizbēgama (faktiem bez likumsakarībām vismaz “vienreizējo gadījumu īpatsvars”); apskatu, kurā lielākā daļa vadošo etalonu “nezinu” nedod punktus; un četru modeļu gadījuma pētījumu, kur punktu noteikumu ierakstīšana uzvednē (“atklāta vērtēšana”) ļauj halucinācijas mazinošai metodei uzvarēt tur, kur parastā precizitāte to sodīja.

Kas ir ticams, bet nav pierādīts: ka atklātas vērtēšanas shēmas, kļūstot par galveno etalonu standartu, būtiski samazinātu halucinācijas reāli izmantotos modeļos. Atbalstošais eksperiments ir neliels un skaidri nekontrolēts.

Ko tas neparāda: ka halucinācijas ir neizbēgamas (darbs argumentē pretējo); ka vērtēšana ir vienīgais cēlonis; ka halucinācijas var vienkārši pilnībā izskaust; ka gadījuma pētījums ir četru modeļu savstarpējs rangs.

Galvenie ierobežojumi: apzināti vienkāršots pareizi/nepareizi/“nezinu” modelis (paši autori to sauc par “viltus trihotomiju”); neliels, atlasīts etalonu pārskats (desmit novērtējumi); nekontrolēts

gadījuma pētījums (četri modeļi, viena metode, viens tests, noklusētie iestatījumi); un OpenAI vadīts arguments par to, kā nozarei vajadzētu vērtēt modeļus.

Cik lielai pārliecībai vajadzētu būt vispārīgam lasītājam? Augstai par to, ka halucinācijas nav ne mistiskas, ne pilnīgi neizbēgamas un ka tipiskie etaloni pašlaik atalgo minēšanu. Vidējai par piedāvātā labojuma plašo efektivitāti — tagad ir reāls principa pierādījums, bet vēl ne demonstrācija lielā mērogā un daudzās situācijās.

Avoti

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