The Clean Paper

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

Miks keelemudelid hallutsineerivad — ja miks meie hindamisviis seda alal hoiab

Enesekindlad valeväited, mida nimetame „hallutsinatsioonideks", pole müstiline rike: osa neist on treeningu statistiline kõrvalsaadus ning need püsivad, sest levinud võrdlustestid premeerivad enesekindlat pakkumist ausa „ma ei tea" asemel. Nelja tippmudeli juhtumiuuring näitab, et punktireeglite kirjutamine viiba sisse („avatud hindamisreeglid") pöörab selle stiimuli ümber.

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

Miks mudelid pakuvad — ja miks me õpetasime neid seda tegema

Küsi suurelt keelemudelilt võõra inimese sünnipäeva ja ta võib vastata „7. märts" sama rahuliku kindlusega, nagu loeks kuupäeva kaardilt — ning eksida kolm korda järjest kolme erineva kuupäevaga. Autorid toovad just sellise näite: tippmudelitelt küsitakse lihtsat faktiküsimust — inimese sünnipäeva või mõne vähetuntud lühendi tähendust — ja igaüks leiutab enesekindlalt erineva vastuse, ükski neist õige. Tööstus nimetab seda *hallutsinatsiooniks*, mis paneb selle kõlama tajuhäirena. Töö esimene samm on võtta sellest müstika välja.

Alustame sellest, kuidas mudel ehitatakse. Esimeses ja suurimas treeninguetapis õpib see sisuliselt tohutut tekstihulka lugedes, milline ladus keel välja näeb. Võtame nüüd fakti, mille taga pole mustrit — ühe konkreetse inimese sünnipäeva. Kui see kuupäev esines treeningtekstis ühe korra või mitte kunagi, pole mustriõppijal millestki kinni võtta: mudeli vaatepunktist on vastus meelevaldne. Autorid teevad selle täpseks, laenates vana idee Alan Turingilt, küll teisest probleemist: kui iga viies sünnipäev esineb andmetes ainult ühe korra, peaks mudel eeldatavasti vähemalt iga viienda puhul eksima — mitte seepärast, et mudel on katki, vaid seepärast, et õppimiseks polnud midagi. (Sama loogika järgi eksivad mudelid riigi pealinna puhul peaaegu mitte kunagi: neid esineb pidevalt.) Autorid väidavad hoolikalt, et tõese väite eristamine usutavast valest on ise raske probleem ning *ainult* tõeste väidete genereerimine on vähemalt sama raske. Teatud veapõrand on paratamatult sees.

Alusidee: kuidas lugeda seda, mida sa pole veel näinud

See põhineb päriselt nutikal ideel — vanemal kui keelemudelid ja väärt lähemalt tundmaõppimist.

Alustame kotitäiest värvilistest pallidest. Sa ei tea, mitu värvi kotis on. Tõmbad ükshaaval 100 palli ja loendad: punaseid 40, siniseid 25, rohelisi 15, kollaseid 5, lillasid 3, oranže 2 — ning siis **kümme erinevat värvi, millest igaüks ilmub täpselt ühe korra.**

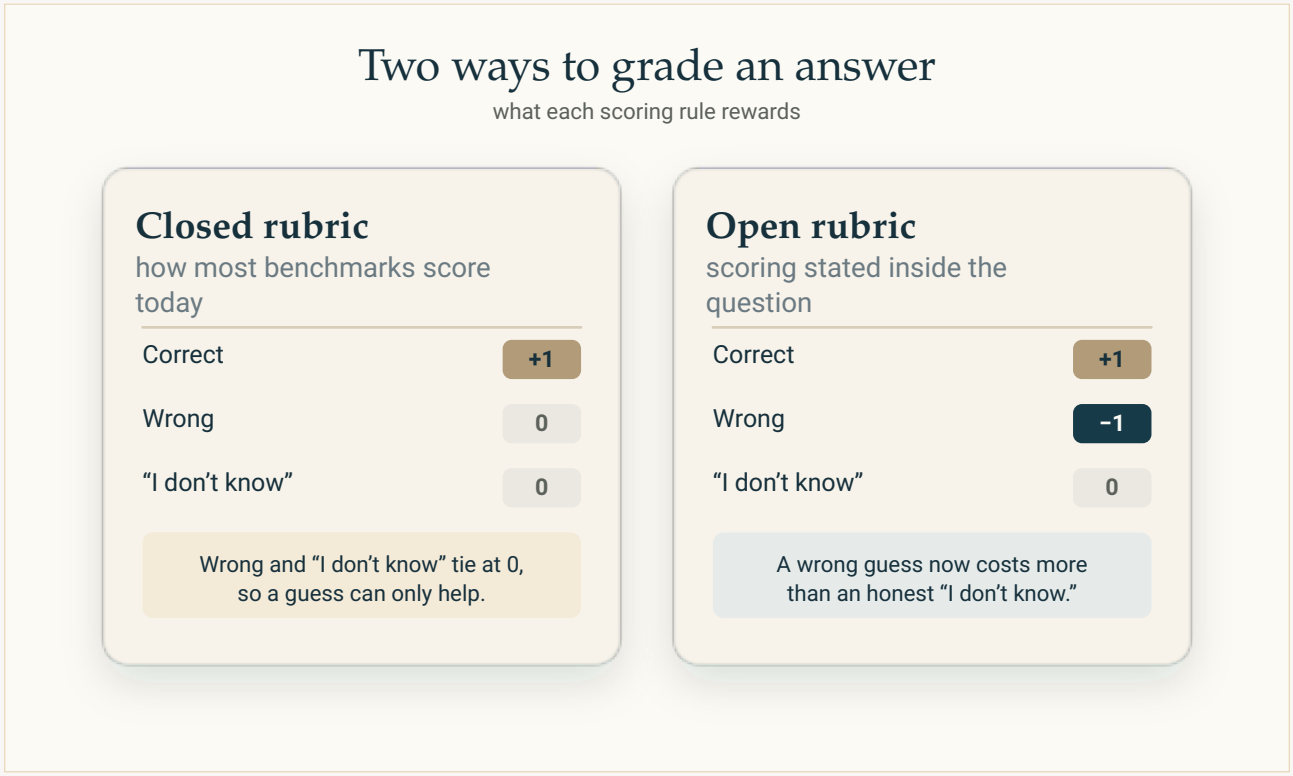
Nüüd küsimus, millega Turing tegelikult hoopis teise probleemi juures kokku puutus: *kui suur on tõenäosus, et järgmine pall on värvi, mida sa pole kordagi näinud?* Seda, mida sa pole tõmmanud, ei saa otse kokku lugeda — aga sa **saad** lugeda värve, mida oled näinud täpselt ühe korra, niinimetatud *singletons*. Võte nimega **Good–Turingi hindamine** ütleb, et ühe korra nähtud vaatluste osakaal hindab tõenäosusmassi, mis peitub veel nägemata värvides. Saja tõmbe hulgas oli kümme ühekorra-värvi, seega on järgmise täiesti uue värvi tõenäosus umbes **10 / 100 = 10%.**

Need ühe korra nähtud värvid ei ole *vead*. Need mõõdavad su teadmatust: kui paljud värvid ilmuvad vaid korra, ütleb valim sulle, et maailmas on veel rohkem sellist, mida sa lihtsalt pole tõmmanud.

Asendame nüüd värvid sünnipäevadega ja koti mudeli treeningtekstiga. Oletame, et nähtud sünnipäevadest **üks viiest esineb täpselt ühe korra.** Sama võte: umbes viiendik tõenäosusmassist asub sünnipäevades, mida mudel pole sisuliselt näinud — ja sünnipäeval pole mustrit, millele toetuda (kellegi sünnipäeva ei saa *välja arvutada*). Ühe korra nähtud või täiesti nägemata kuupäev on münt, mida mudel ei oska kaaluda, ning ta eksib umbes **ühel juhul viiest**. Ükski nutikus ei paranda seda: õppimiseks polnud midagi.

> See on kogu argument miniatuuris: singleton-määr mõõdab, kui suur osa maailmast on *nendest andmetest* õppimatu, ning sellest saab vigade **alumine piir**. Samuti selgitab see, miks mudel pealinnaga peaaegu kunagi mööda ei pane — *Pariis* esineb pidevalt, selle singleton-määr on peaaegu null ja õppimiseks on küllaga materjali.

See selgitab, kust hallutsinatsioonid tulevad. See ei selgita, miks need *püsivad* — miks mudelid pärast kogu hilisemat treeningut, mille eesmärk on muuta nad kasulikuks ja ausaks, ikka bluffivad selle asemel, et ebakindlust tunnistada. Siin on töö analoogia peaaegu ebamugavalt täpne. Kujuta ette õpilast eksamil, kes vastust ei tea. Kui tühi vastus annab null punkti ja pakkumine võib anda ühe, on hinde maksimeerimiseks õige käik pakkuda — enesekindlalt, konkreetselt, mitte kunagi „ma pole kindel". Õpilased õpivad seda. Nagu selgub, mudelid samuti — sest me hindame neid samamoodi. Autorid vaatasid läbi võrdlustestid, mille nimel valdkond päriselt võistleb, edetabelid, kuhu mudelid optimeeritakse, ning leidsid, et peaaegu kõik annavad „ma ei tea" eest täpselt sama tulemuse kui vale vastuse eest: null. Sellise reegli all võidab alati pakkuv mudel muidu identsest mudelist, mis märgib ausalt oma ebakindluse. Üsna otseses mõttes hindame me neid hallutsineerima.



Võrdlustestid võivad muuta pakkumise ratsionaalseks. Enamiku testide kasutatava hindamise korral (vasakul) saavad vale vastus ja aus „ma ei tea" mõlemad null punkti — seega saab pakkumine ainult aidata. Kui reeglid öelda küsimuses välja ja vale vastuse eest anda karistus (paremal), võib ebakindluse korral vastamata jätmine olla parem käik. See muudab seda, mida test premeerib; iseenesest ei lahenda see hallutsinatsioone.

Just see osa tasub meelde jätta, sest see läheb vastu tavapärasele pealkirjale. Hallutsinatsiooni müüakse sageli kui tehnoloogia vältimatut, peaaegu müstilist piiri. Töö vaidleb mõlemale poolele vastu. Eeltreeningu veapõrand ei ole müsteerium — see on tavaline statistiline viga, mida masinõpe

on aastakümneid mõistnud. Ja püsimine ei ole vältimatu — vähemalt osaliselt on see meie endi ehitatud stiimul, mida saaks muuta. Süsteem, mis ebakindluse korral lihtsalt keeldub vastamast, ei hallutsineeriks üldse; kasutuses olevad mudelid ei käitu nii, sest meie edetabelid karistavad keeldumist.

Mida autorid tegid

Tööl on kolm osa. Esiteks matemaatiline argument, et osa hallutsinatsioonist on eeltreeningu ajal statistiliselt sunnitud, näidates, et „genereeri ainult kehtiv tekst" on vähemalt sama raske kui binaarne klassifitseerimisprobleem „kas see väide on kehtiv?". Teiseks argument — mida toetab kümne mõjuka võrdlustesti ülevaade — et levinud täpsuspõhised mõõdikud premeerivad pakkumist vastamata jätmise asemel. Kolmandaks pakutud parandus ja seda testiv juhtumiuuring: **avatud hindamisreeglitega** (*open-rubric*) hindamine, kus punktireegel on kirjas küsimuses endas (näiteks „õige vastus annab 1, vale −1, seega jäta vastamata, kui oled vähem kui 50% kindel"), nii et mudel teab, millal ausust premeeritakse. Nad proovivad seda neljal tippmudelil — Google Gemini 3 Pro, OpenAI GPT-5, xAI Grok 4 ja Anthropic Claude Opus 4.5 — kasutades SimpleQA 4326 faktiküsimust. Autorid rõhutavad, et juhtumiuuring on illustratiivne, „mitte kontrollitud mudelitevaheline hindamine" (vaikesätted, häälestamist ega kulude normaliseerimist pole).

Mida nad leidsid

- **Eeltreening sunnib osa vigu.** Määra, millega mudel toodab enesekindlaid valeväiteid, piirab altpoolt (ligikaudu kahekordselt) temast ehitatud parima „kas see väide on kehtiv?" klassifikaatori veamäär. Faktide puhul, millel puudub õpitav muster, on see põrand vähemalt *singleton-määr* — nende faktide osakaal, mis esinevad treeningus täpselt ühe korra. Mõni hallutsinatsioon on vältimatu isegi täiesti puhaste andmete korral.
- **Hindamine premeerib pakkumist — konkreetselt.** Tavalise õige/vale punktiarvestuse puhul on optimaalne strateegia mitte kunagi vastamata jätta ning autorite ülevaade leiab, et suur enamus populaarseid võrdlusteste hindab „ma ei tea" lihtsalt valeks. Ilmekas näide nende endi poolelt: SimpleQA testis *soosib* toortäpsus veidi OpenAI o4-mini mudelit — mis vastab peaaegu kõigele ja eksib rohkem kui kolmel neljandikul juhtudest — GPT-5-mini ees, mis teeb palju vähem vigu, sest jätab ebakindluse korral vastamata. Hoolimatum mudel näeb edetabelis parem välja.
- **Avatud hindamisreeglid pööravad stiimuli ümber (nende juhtumiuuringus).** Nad testivad lihtsat hallutsinatsiooni vähendamise võtet: las mudel vastab kaks korda ning jätab vastamata, kui vastused ei ühti. Tavapärase täpsuse puhul vähendab võte vigu, *aga vähendab ka täpsust* — seega mõõdik heidutab selle kasutamist. Avatud hindamisreeglitega tuleb sama võte kõigil neljal mudelil eri karistusmäärade puhul paremini välja; ning GPT-5-mini — mida toortäpsus karistas ebakindluse korral vastamata jätmise eest — edestab o4-mini mudelit, kui punktireegel on avalikult kirjas (n = 4326 küsimust mudeli kohta).

Mida see tõenäoliselt tähendab

Hallutsinatsiooni vähendamine ei ole peamiselt küsimus selles, kuidas leiutada veel rohkem hallutsinatsioonispetsiifilisi teste. Küsimus on selles, kuidas muuta levinud võrdlustestide ebakindluse hindamist nii, et „ma ei tea" tunnistamine poleks enam karistatav. Kuni edetabel ei muutu, maksab hallutsinatsiooni vähendamine mudelile täpsuspunkte ja jääb seetõttu ebasoovitavaks — seepärast nimetavad autorid probleemi „sotsio-tehniliseks": osaliselt on vaja paremat mõõdikut, osaliselt peab veenma mõjukaid edetabeleid seda kasutusele võtma.

Mida see *ei* tõesta

- See **ei** näita, et avatud hindamisreeglid parandavad hallutsinatsiooni päris kasutuses. Toetav katse on väike ja teadlikult **kontrollimata** juhtumiuuring — neli vaikesätetel mudelit, üks valitud leevendusvõte, üks faktiküsimuste test — mille eesmärk on näidata *stiimuli ümberpöördumist*, mitte mudelite järjestamist ega üldise tõhususe tõestamist.
- See **ei** väida, et hindamine on *ainus* põhjus. Treeningandmete vead, päriselt rasked probleemid ja tundmatud viibad jäävad eraldi allikateks.
- See **ei** toeta populaarset väidet, et hallutsinatsioonid on vältimatud. Autorid väidavad vastupidist: süsteem, mis vastaks ainult kontrollitavatele küsimustele ja muidu ütleks „ma ei tea", ei hallutsineeriks kunagi.
- See **ei** kaota eeltreeningu veapõrandat — see selgitab ja piirab seda ning piir puudutab enesekindlaid faktivigu, mitte mudeli kogu käitumist.
- See **ei** näita, et avatud hindamisreeglitest *üksi piisab*. Need muudavad seda, mida hindamine premeerib; need ei asenda infootsingut, tööriistade kasutamist ega paremini kalibreeritud mudeleid.

Kui tugevad on tõendid?

- Tuum on **matemaatika** — formaalsed alumised piirid, mitte mõõtmised. Teoreetilise argumendina on see oma eelduste piires tugev.
- See toetub probleemi **teadlikult lihtsustatud mudelitele**; autorid ise märgivad „vale trihhotoomia" probleemi, kui iga vastus liigitatakse õigeks, valeks või „ma ei tea", ning idealiseeritud „meelevaldsete faktide" keskkonda, mida kasutatakse puhtaima piiri jaoks.
- Võrdlustestide ülevaade on **väike kureeritud valim** — kümme mõjukat hindamist, mitte ammendav audit.
- Juhtumiuuring on **päris, kuid piiratud**: neli tippmudelit, üks leevendusvõte, ainult SimpleQA, vaikesätted, selgesõnaliselt „mitte kontrollitud hindamine". See on stiimuliargumendi põhimõtte tõestus, mitte võrdlustesti tulemus.
- Tasub nimetada vaatepunkti: neli autorit kolmest töötavad või on töötanud **OpenAI-s** ning töö väidab, et valdkond peaks muutma mudelite hindamist. See on huvitatud osapoole hästi argumenteeritud seisukoht, mitte neutraalne välisülevaade — seda tuleb kaaluda, mitte kõrvale heita.

(Töö kasuks räägib, et kriitika suunatakse sama valmisolekuga oma mudelite o4-mini ja GPT-5-mini kui teiste pihta.)

Miks see oluline on

See raamib ümber tugevalt ülesköetud probleemi. „Hallutsinatsiooni" müüakse kas õõvastava veana või nihutamatu seinana; see töö muudab selle tavaliseks ja osaliselt enda tekitatuks — statistiline põrand, millest saame päriselt aru, ning selle peal meie enda valitud stiimul. Laiem õppetund on vaiksem ja kasulikum: järgmine samm töökindluses võib sõltuda sama palju sellest, *mida me mõõdame,* kui sellest, mida me ehitame.

Lühidalt

Keelemudelite enesekindlad valevastused tulevad kahest kohast. Esimene on statistiline: kui faktil pole õpitavat mustrit, eksib keelt jäljendama treenitud mudel mõnikord ning selle põranda suurust saab hinnata, näiteks selle järgi, kui palju fakte ilmub treeningus ainult ühe korra. Teine on stiimulites: peaaegu kõik võrdlustestid, mille järgi mudeleid järjestatakse, annavad „ma ei tea" eest sama tulemuse kui vale vastuse eest, seega võidab pakkumine alati — kuni selleni, et kolmveerandil juhtudest eksiv mudel võib saada kõrgema skoori kui ausam mudel, mis jätab ebakindluse korral vastamata. Autorite ettepanek pole uus hallutsinatsioonitest, vaid „avatud hindamisreeglid": punktiarvestus öeldakse küsimuses välja. Nelja tippmudeli juhtumiuuringus pöörab see stiimuli ümber, nii et hallutsinatsiooni vähendavat meetodit premeeritakse, mitte ei karistata. Tegemist on teooriat ja ülevaadet ühendava tööga, millel on väike ja selgelt kontrollimata katse ning mis on eelretsenseeritud ajakirjas *Nature*; lahendus on paljulubav, kuid selle laiaulatuslikku toimimist pole veel näidatud ning autorite järgi pole hallutsinatsioonid ei müstilised ega rangelt vältimatud.

Ilustamata kontroll

Mida töö näitab: Matemaatiline alumine piir, mille tõttu osa hallutsinatsioonist on eeltreeningus sunnitud (mustrita faktide puhul vähemalt „singleton-määr"); ülevaade, mille järgi enamik juhtivaid võrdlusteste ei anna „ma ei tea" eest krediiti; ning nelja mudeli juhtumiuuring, kus punktireegli viibale lisamine („avatud hindamisreeglid") paneb hallutsinatsiooni vähendava meetodi võitma olukorras, kus tavaline täpsus seda karistas.

Mis on usutav, kuid tõestamata: Et avatud hindamisreeglite viimine peamistesse võrdlustestidesse vähendaks kasutuses olevate mudelite hallutsinatsioone sisuliselt. Toetav katse on väike ja selgesõnaliselt kontrollimata.

Mida see ei näita: Et hallutsinatsioonid on vältimatud (töö väidab vastupidist); et hindamine on ai-

nus põhjus; et hallutsinatsiooni saab täielikult kõrvaldada; et juhtumiuuring järjestab neli mudelit omavahel.

Peamised piirangud: Teadlikult lihtsustatud õige/vale/„ma ei tea” mudel (autorid nimetavad seda „valeks trihhotoomiaks”); väike kureeritud võrdlustestide ülevaade (kümme hindamist); kontrollimata juhtumiuuring (neli mudelit, üks leevendus, üks test, vaikesätted); ning OpenAI-ga tugevalt seotud autorite argument selle kohta, kuidas valdkond peaks mudeleid hindama.

Kui palju võiks tavalugeja seda usaldada? Kõrge kindlus, et hallutsinatsioon pole ei müstiline ega rangelt vältimatu ning et levinud võrdlustestid premeerivad praegu pakkumist. Mõõdukas kindlus, et pakutud parandus aitab — sellel on nüüd päris põhimõtte tõestus, kuid mitte veel demonstratsioon, et see toimib laialt ja suures mastaabis.

Allikad

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