The Clean Paper

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

Miksi kielimallit hallusinoivat — ja miksi pisteytystapamme pitää ilmiötä yllä

Itsevarmat väärät väitteet, joita kutsumme ”hallusinaatioiksi”, eivät ole mystinen toimintahäiriö: osa on koulutuksen tilastollinen sivutuote, ja ne säilyvät, koska tavalliset vertailut palkitsevat itsevarman arvauksen rehellisen ”en tiedä” -vastauksen sijaan. Neljällä frontier-mallilla tehty tapaustutkimus osoittaa, että pisteytyssääntöjen kertominen itse kehotteessa (”open rubrics”) kääntää tämän kannustimen.

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

Miksi mallit arvaavat — ja miksi opetimme ne tekemään niin

Kysy suurelta kielimallilta tuntemattoman ihmisen syntymäpäivää, ja se saattaa vastata ”7. maaliskuuta” samalla vakaalla varmuudella kuin lukisi tiedon kortista — ja olla väärässä kolme kertaa peräkkäin kolmella eri päivämäärällä. Kirjoittajat antavat juuri tällaisia esimerkkejä: johtavilta malleilta kysytään yksinkertainen faktakysymys — henkilön syntymäpäivä tai harvinaisen lyhenteen merkitys — ja jokainen keksii itsevarmasti erilaisen vastauksen, joista yksikään ei ole oikea. Ala kutsuu tätä *hallusinaatioksi*, mikä saa ilmiön kuulostamaan havaintovirheeltä. Paperin ensimmäinen askel on poistaa siitä mystiikka.

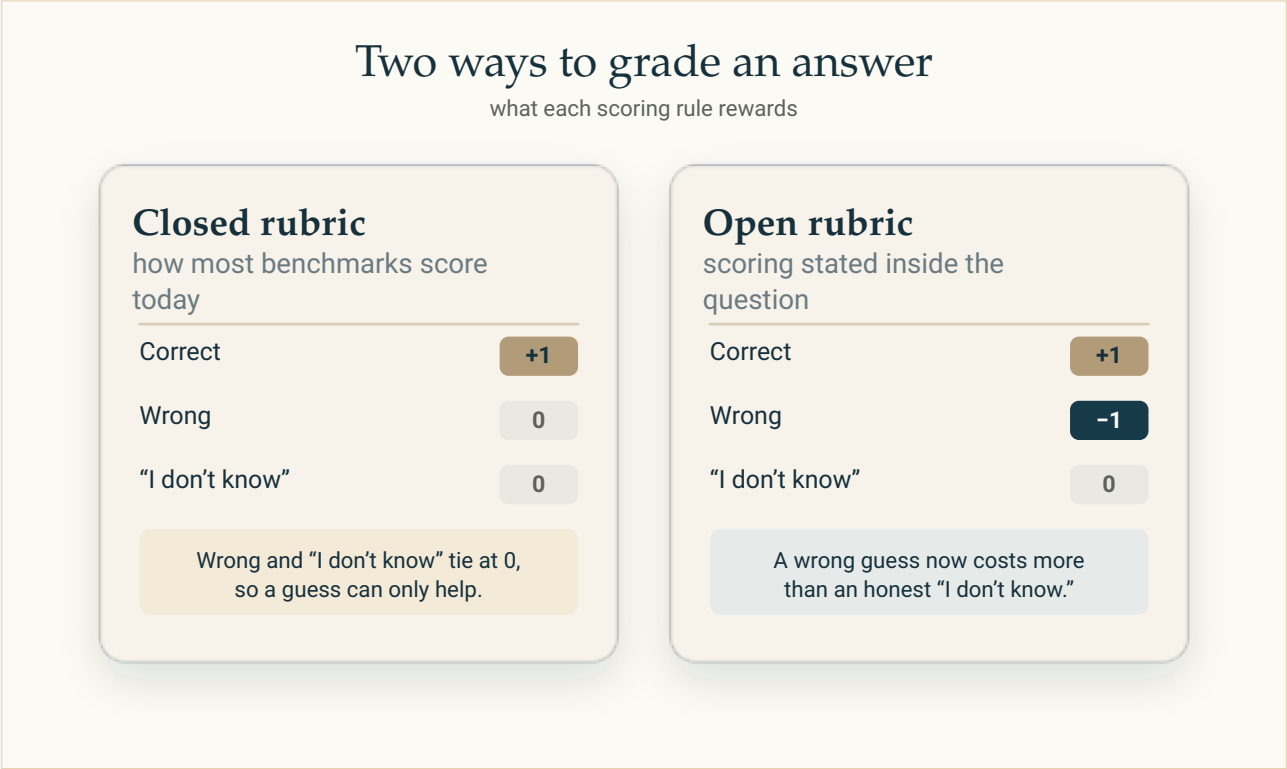
Aloitetaan siitä, miten malli rakennetaan. Ensimmäisessä ja suurimmassa koulutusvaiheessaan se oppii valtavasta tekstimäärästä käytännössä, miltä sujuva kieli näyttää. Otetaan sitten fakta, jonka takana ei ole mitään opittavaa kaavaa — yhden tietyn ihmisen syntymäpäivä. Jos päivämäärä esiintyi koulutustekstissä kerran tai ei lainkaan, kaavoja oppivalla järjestelmällä ei ole mitään mihin tarttua: mallin näkökulmasta vastaus on mielivaltainen. Kirjoittajat täsmentävät tämän lainaamalla vanhaa ideaa (Alan Turingilta, toiseen ongelmaan): jos joka viides syntymäpäivä esiintyy aineistossa vain kerran, mallin pitäisi odottaa erehtyvän vähintään joka viidennessä — ei siksi, että se olisi rikki, vaan koska opittavaa ei koskaan ollut. (Samalla logiikalla mallit erehtyvät erittäin harvoin valtioiden pääkaupungeista: ne esiintyvät jatkuvasti.) Kirjoittajat argumentoivat huolellisesti, että toden väitteen erottaminen uskottavasta väärästä on itsessään vaikea ongelma ja että *vain* tosien väitteiden tuottaminen on vähintään yhtä vaikeaa. Virheille syntyy alaraja.

Taustalla oleva idea: miten lasketaan se, mitä ei ole vielä nähty

Tämän alla on aidosti nokkela idea — kielimalleja vanhempi ja tutustumisen arvoinen.
Aloita pussillisesta värillisiä palloja. Et tiedä, kuinka monta väriä pussissa on. Vedät 100 palloa yksi kerrallaan ja lasket ne: punaista 40, sinistä 25, vihreää 15, keltaista 5, violettia 3, oranssia 2 — ja sitten **kymmenen eri väriä, joista kukin esiintyy tasan kerran.**
Kysymys, jonka Turing oikeasti kohtasi aivan toisessa ongelmassa, oli: *mikä on todennäköisyys, että seuraava pallo on väriä, jota et ole vielä nähnyt lainkaan?* Et voi laskea sellaista, mitä et ole koskaan nostanut — mutta **voit** laskea värit, jotka olet nähnyt tasan kerran, eli ”singletonit”.
Good–Turing-estimaation idea on, että singletonien osuus nostoista arvioi todennäköisyysmassaa, joka piilee vielä näkemättömissä väreissä. Sadasta nostosta kymmenen oli kerran nähtyä väriä, joten seuraavan täysin uuden värin todennäköisyys on noin **10 / 100 = 10%.**
Kerran nähdyt värit eivät ole *virheitä*. Ne mittaavat omaa tietämättömyyttäsi: kun moni väri esiintyy kerran, otos kertoo, että maailmassa on lisää sellaista, jota et vain ole vielä saanut näkyviin.
Vaihda nyt värit syntymäpäiviksi ja pussi mallin koulutustekstiksi. Oletetaan, että mallin näkemistä syntymäpäivistä **joka viides esiintyy tasan kerran.** Sama temppu: noin viidennes todennäköisyysmassasta kuuluu syntymäpäiviin, joita malli ei käytännössä ole nähnyt — eikä syntymäpäivässä ole kaavaa, johon nojata (ihmisen syntymäpäivää ei voi *päätellä*). Kerran nähty tai kokonaan näkemätön päivämäärä on siis kuin kolikko, jota malli ei osaa painottaa, ja se erehtyy noin **yhdessä viidestä**. Mikään nokkeluus ei korjaa tätä: opittavaa ei ollut.

> Tämä on koko argumentti pienoiskoossa: singleton-osuus mittaa, kuinka suuri osa maailmasta on *tästä aineistosta* oppimattomissa, ja siitä tulee virheiden **alaraja**. Siksi malli myös erehtyy pääkaupungista hyvin harvoin — *Paris* esiintyy jatkuvasti, sen singleton-osuus on lähellä nollaa ja opittavaa dataa on paljon.

Tämä selittää, mistä hallusinaatioita tulee. Se ei selitä, miksi ne *säilyvät* — miksi mallit kaiken myöhemmän hyödyllisyyteen ja rehellisyyteen tähtäävän koulutuksen jälkeen edelleen bluffaavat sen sijaan, että myöntäisivät epävarmuuden. Tässä paperin vertaus on lähes epämukavan osuva. Kuvittele kokeessa oleva opiskelija, joka ei tiedä vastausta. Jos tyhjästä saa nolla pistettä ja arvauksesta voi saada yhden, arvosanan maksimoiva strategia on arvata — itsevarmasti ja täsmällisesti, ei koskaan ”en ole varma”. Opiskelijat oppivat tämän. Niin näköjään mallitkin, koska pisteytämme niitä samalla tavalla. Kirjoittajat kävivät läpi vertailut, joilla ala oikeasti kilpailee ja joiden tulostaulukoita varten malleja viritetään, ja havaitsivat lähes kaikkien antavan ”en tiedä” -vastaukselle täsmälleen saman pistemäärän kuin väärälle vastaukselle: nolla. Tällaisella säännöllä aina arvaava malli päihittää muuten identtisen mallin, joka ilmoittaa epävarmuutensa rehellisesti. Melko kirjaimellisesti pisteytämme mallit arvaamaan.



Vertailut voivat tehdä arvaamisesta rationaalista. Useimpien vertailujen tavallisessa pisteytyksessä (vasemmalla) väärä vastaus ja rehellinen ”en tiedä” saavat molemmat nolla pistettä — joten arvaamisesta voi vain hyötyä. Kun säännöt kerrotaan itse kysymyksessä ja väärästä vastauksesta tulee rangaistus (oikealla), vastaamatta jättäminen epävarmana voi muuttua paremmaksi ratkaisuksi. Tämä muuttaa sitä, mitä testi palkitsee; se ei yksin ratkaise hallusinaatiota.

Original diagram — The Clean Paper CC BY 4.0

Tämä osa kannattaa muistaa, koska se kulkee tavallista otsikkoa vastaan. Hallusinaatio esitetään

usein väistämättömänä, lähes mystisenä teknologian rajana. Paperi kiistää molemmat kuvaukset. Esikoulutuksen virhelattia ei ole mysteeri — se on tavallista tilastollista virhettä, sellaista jota koneoppimisessa on ymmärretty vuosikymmeniä. Eikä ilmiön jatkuminen ole väistämätöntä — osittain se johtuu kannustimesta, jonka rakensimme itse ja jonka voisimme muuttaa. Järjestelmä, joka epävarmana yksinkertaisesti kieltäytyisi vastaamasta, ei hallusinoisi lainkaan; käytössä olevat mallit eivät käyttäydy näin, koska tulostaulumme rankaisevat kieltäytymisestä.

Mitä kirjoittajat tekivät

Paperissa on kolme osaa. Ensimmäinen on matemaattinen argumentti siitä, että esikoulutus pakottaa jonkin verran hallusinaatiota tilastollisesti: ”tuota vain kelvollista tekstiä” osoitetaan vähintään yhtä vaikeaksi kuin binäärinen luokitteluongelma ”onko tämä väite kelvollinen?”. Toinen on kymmenen vaikutusvaltaisen vertailun katsaukseen tukeutuva argumentti siitä, että tavalliset tarkkuusmittarit palkitsevat arvaamista vastaamatta jättämisen sijaan. Kolmas on ehdotettu korjaus ja sitä testaava tapaustutkimus: **open-rubric**-arvioinnit, joissa pisteytys kerrotaan itse kysymyksessä (esimerkiksi ”oikeasta vastauksesta 1 piste, väärästä −1, joten jätä vastaamatta jos olet alle 50% varma”), jotta malli tietää, milloin rehellisyys palkitaan. He kokeilevat tätä neljällä frontier-mallilla — Googlen Gemini 3 Prolla, OpenAI:n GPT-5:llä, xAI:n Grok 4:llä ja Anthropicin Claude Opus 4.5:llä — käyttäen SimpleQA-testin 4,326 faktakysymystä. Kirjoittajat sanovat suoraan, että tapaustutkimus on havainnollistava, ”ei kontrolloitu mallien välinen arviointi” (oletusasetukset, ei viritystä, ei kustannusten normalisointia).

Mitä he löysivät

- **Esikoulutus pakottaa jonkin verran virhettä.** Mallin tuottamien itsevarmojen väärien väitteiden määrällä on alaraja, joka määräytyy siitä rakennetun parhaan ”onko tämä väite kelvollinen?” -luokittimen virheprosentista (suunnilleen kaksinkertaisena). Faktoille, joissa ei ole opittavaa kaavaa, tämä alaraja on vähintään *singleton-osuus* — niiden faktojen osuus, jotka esiintyvät koulutuksessa tasan kerran. Jonkin verran hallusinaatiota on väistämätöntä jopa täysin puhtaalla datalla.
- **Pisteytys palkitsee arvaamista — konkreettisesti.** Tavallisessa oikein/väärin-pisteytyksessä koskaan vastaamatta jättämätön strategia on optimaalinen, ja kirjoittajien katsauksessa suuri enemmistö suosituista vertailuista pisteyttää ”en tiedä” -vastauksen yksinkertaisesti vääräksi. Heidän omalta puoleltaan havainnollinen esimerkki: SimpleQA-testissä raakaversio tarkkuudesta hieman *suosii* OpenAI:n o4-miniä — joka vastaa lähes kaikkeen ja on väärässä yli kolmessa neljäsosassa — GPT-5-minin sijaan, joka tekee paljon vähemmän virheitä, koska jättää epävarmana vastaamatta. Holtittomampi malli näyttää tulostaululla paremmalta.
- **Open rubrics kääntää kannustimen (heidän tapaustutkimuksessaan).** He testaavat yksinkertaista hallusinaation lievennystä: malli vastaa kahdesti ja jättää vastaamatta, jos vastaukset eroavat. Tavallisella tarkkuusmittarilla menetelmä vähentää virheitä *mutta vähentää myös tarkkuutta* — joten mittari rankaisee sen käyttämisestä. Avoimella pisteytyksellä sama lievennys

voittaa kaikilla neljällä mallilla useilla eri virherangaistuksilla; ja GPT-5-mini — jota raakaversio tarkkuudesta rankaisi epävarmana pidättäytymisestä — nousee o4-minin edelle, kun pisteytys kerrotaan avoimesti (n = 4,326 kysymystä mallia kohti).

Mitä tämä todennäköisesti tarkoittaa

Hallusinaation vähentäminen ei ehkä ensisijaisesti tarvitse lisää juuri hallusinaatioita varten rakennettuja testejä. Tarvitaan muutosta siihen, miten keskeiset vertailut pisteyttävät epävarmuuden, jotta ”en tiedä” ei enää ole rangaistava vastaus. Niin kauan kuin tulostaulu ei muutu, hallusinaatioiden vähentäminen maksaa malleille tarkkuuspisteitä ja sitä siis edelleen lannistetaan — siksi kirjoittajat kutsuvat ongelmaa ”sosio-tekniseksi”: osa on parempi mittari, osa vaikutusvaltaisten tulostaulujen saaminen käyttämään sitä.

Mitä tämä *ei* osoita

- Se **ei** osoita, että open rubrics korjaa hallusinaatiot käytännön maailmassa. Tukeva koe on pieni ja tarkoituksella **kontrolloimaton** tapaustutkimus — neljä mallia oletusasetuksilla, yksi valittu lievennys ja yksi faktakysymystesti — jonka tarkoitus on osoittaa *kannustimen kääntyminen*, ei asettaa malleja paremmuusjärjestykseen tai todistaa yleistä tehoa.
- Se **ei** väitä pisteytyksen olevan *ainoa* syy. Koulutusdatan virheet, aidosti vaikeat ongelmat ja oudot kehotteet ovat edelleen erillisiä lähteitä.
- Se **ei** tue suosittua väitettä, että hallusinaatiot olisivat väistämättömiä. Kirjoittajat väittävät päinvastoin: järjestelmä, joka vastaisi vain tarkistettaviin kysymyksiin ja sanoisi muuten ”en tiedä”, ei hallusinoisi koskaan.
- Se **ei** poista esikoulutuksen virhelattiaa — se selittää ja rajaa sen, ja raja koskee itsevarmoja faktavirheitä, ei kaikkea mallin käyttäytymistä.
- Se **ei** osoita, että open rubrics olisi yksinään *riittävä*. Se muuttaa arvioinnin palkitsemia asioita; se ei korvaa tiedonhakua, työkalujen käyttöä tai paremmin kalibroituja malleja.

Kuinka vahvaa näyttö on?

- Ydin on **matematiikkaa** — muodollisia alarajoja, ei mittauksia. Teoreettisena argumenttina se on pätevä omilla oletuksillaan.
- Se nojaa ongelman **tarkoituksella yksinkertaistettuihin malleihin**; kirjoittajat itse nostavat esiin ”väärän trikotomian”, jossa jokainen vastaus käsitellään oikeana, vääränä tai ”en tiedä” -vastauksena, sekä puhtainta alarajaa varten käytetyn idealisoidun ”mielivaltaisten faktojen” asetelman.
- Vertailukatsaus on **pieni, kuratoitu otos** — kymmenen vaikutusvaltaista arviointia, ei kattava auditointi.

- Tapaustutkimus on **todellinen mutta rajallinen**: neljä frontier-mallia, yksi lievennys, vain SimpleQA, oletusasetukset ja nimenomaisesti ”ei kontrolloitu arviointi”. Se on ensimmäinen käytännön osoitus kannustinargumentista, ei vertailutestin tulos.
- Näkökulma on myös syytä nimetä: kolme neljästä kirjoittajasta työskentelee tai on työskennellyt **OpenAI:lla**, ja paperi argumentoi sen puolesta, että alan pitäisi muuttaa mallien arviointitapaa. Kyse on hyvin perustellusta kannasta asianosaiselta, ei neutraalista ulkopuolisesta katsauksesta — asia punnittavaksi, ei hylättäväksi. (Paperin eduksi se kohdistaa kritiikin yhtä valmiisti omiin o4-mini- ja GPT-5-mini-malleihinsa kuin muihin.)

Miksi tämä on tärkeää

Se kehystää voimakkaasti hypetetyn ongelman uudelleen. ”Hallusinaatio” myydään usein joko aavemaisena vikana tai liikkumattomana seinänä; tämä paperi tekee siitä tavallisen ja osin itse aiheutetun — ymmärrettävissä olevan tilastollisen lattian, jonka päällä on itse valitsemamme kannustin. Laajempi opetus on hiljaisempi ja hyödyllisempi: luotettavuuden seuraavat parannukset voivat riippua yhtä paljon *siitä mitä mittaamme* kuin siitä, mitä rakennamme.

Tiivistelmä

Kielimallien itsevarmat väärät vastaukset tulevat kahdesta paikasta. Ensimmäinen on tilastollinen: kun faktassa ei ole opittavaa kaavaa, kieltä jäljittelemään koulutettu malli erehtyy joskus, ja tämän virhelattian voi arvioida esimerkiksi siitä, kuinka moni fakta esiintyy koulutuksessa vain kerran. Toinen on kannustin: lähes jokainen vertailu, jolla malleja asetetaan paremmuusjärjestykseen, antaa ”en tiedä” -vastaukselle saman pistemäärän kuin väärälle vastaukselle, joten arvaaminen voittaa aina — niin pitkälle, että kolme neljäsosaa ajasta väärässä oleva malli voi päihittää rehellisemmän mallin, joka pidättäytyy epävarmana. Kirjoittajien ehdotus ei ole uusi hallusinaatiotesti vaan ”open rubrics”: pisteytys kerrotaan itse kysymyksessä. Neljän frontier-mallin tapaustutkimuksessa se kääntää kannustimen niin, että hallusinaatioita vähentävää menetelmää palkitaan eikä rangaista. Kyse on teoriasta, katsauksesta ja pienestä, nimenomaisesti kontrolloimattomasta kokeesta koostuvasta vertaisarvioidusta *Nature*-paperista; korjaus on lupaava mutta sen laajamittaista toimivuutta ei ole vielä osoitettu, ja kirjoittajien argumentin mukaan hallusinaatiot eivät ole mystisiä eivätkä ehdottoman väistämättömiä.

Realistinen arvio

Mitä paperi osoittaa: Matemaattisen alarajan, jonka vuoksi esikoulutuksessa syntyy jonkin verran hallusinaatiota (kaavattomille faktoille vähintään ”singleton-osuus”); katsauksen, jonka mukaan useimmat johtavat vertailut eivät anna ”en tiedä” -vastauksesta mitään hyvitystä; sekä neljän mallin tapaustutkimuksen, jossa pisteytyksen kertominen kehotteessa (”open rubrics”) saa hallusinaatioita

vähentävän menetelmän voittamaan tilanteessa, jossa tavallinen tarkkuus rankaisi sitä.

Mikä on uskottavaa mutta ei osoitettu: Että keskeisiin vertailuihin sisällytetyt open rubrics -käytännöt vähentäisivät merkittävästi käytössä olevien mallien hallusinaatioita. Tukeva koe on pieni ja nimenomaisesti kontrolloimaton.

Mitä se ei osoita: Että hallusinaatiot olisivat väistämättömiä (paperi argumentoi päinvastoin); että pisteytys olisi ainoa syy; että hallusinaatio voidaan poistaa kokonaan; että tapaustutkimus asettaisi neljä mallia paremmuusjärjestykseen.

Tärkeimmät rajoitukset: Tarkoituksella yksinkertaistettu oikein/väärin/”en tiedä” -malli (kirjoittajat kutsuvat sitä ”vääräksi trikotomiaksi”); pieni kuratoitu vertailukatsaus (kymmenen arviointia); kontrolloimaton tapaustutkimus (neljä mallia, yksi lievennys, yksi testi, oletusasetukset); sekä OpenAI-vetoinen argumentti siitä, miten alan pitäisi arvioida malleja.

Kuinka paljon luottamusta tavallisella lukijalla pitäisi olla? Paljon siihen, ettei hallusinaatio ole mystinen eikä ehdottoman väistämätön ja että valtavirran vertailut tällä hetkellä palkitsevat arvaamista. Kohtalaisesti siihen, että ehdotettu korjaus auttaa — siitä on nyt todellinen käytännön osoitus, mutta ei vielä osoitusta laajasta ja skaalautuvasta toimivuudesta.

Lähteet

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