The Clean Paper

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

Waarom taalmodelle hallusineer — en waarom ons manier van punte gee dit so hou

Die selfversekerde onwaarhede wat ons "hallusinasies" noem, is geen misterieuse fout nie: 'n deel is 'n statistiese byproduk van opleiding, en hulle bly voortbestaan omdat hoofstroommaatstawwe 'n selfversekerde raaiskoot hoër beloon as 'n eerlike "ek weet nie". 'n Gevallestudie op vier voorste modelle wys dat die noem van die puntereëls in die opdrag ("oop rubrieke") daardie aansporing omkeer.

Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso · 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

Waarom modelle raai — en waarom ons hulle dit geleer het

Vra ’n groot taalmodel na die verjaardag van ’n vreemdeling en dit antwoord dalk “7 Maart” — met die kalm sekerheid van iemand wat dit van ’n kaartjie aflees — en is verkeerd, drie keer agtereenvolgens, met drie verskillende datums. Die outeurs gee presies hierdie soort voorbeeld: voorste modelle wat ’n gewone feitevraag gevra is — iemand se verjaardag, of waarvoor ’n obskure akroniem staan — het elkeen met selfvertroue ’n ander antwoord versin, nie een daarvan reg nie. Die bedryf se woord hiervoor is *hallusinasie*, wat na ’n waarnemingsfout klink. Die artikel se eerste skuif is om die misterie daaruit te haal.

Begin by hoe ’n model gebou word. In sy eerste en grootste opleidingsfase leer dit, in wese, hoe vlot taal lyk deur ’n enorme hoeveelheid teks te lees. Neem nou ’n feit sonder ’n patroon daaragter — een bepaalde persoon se verjaardag. As daardie datum een keer in die opleidingsteks verskyn het, of nooit, is daar niks waaraan ’n patroonleerder kan vat nie: die antwoord is, uit die model se oogpunt, arbitrêr. Die outeurs maak dit presies deur ’n ou idee te leen (Alan Turing s’n, vir ’n ander probleem): as een uit vyf verjaardae net een keer in die data opduik, moet ’n mens verwag dat ’n model minstens een uit vyf daarvan verkeerd kry — nie omdat dit stukkend is nie, maar omdat daar nooit iets was om te leer nie. (Volgens dieselfde logika kry modelle ’n land se hoofstad byna nooit verkeerd nie: dié verskyn aanhoudend.) Hulle voer aan, met ’n mate van sorg, dat om ’n ware stelling van ’n geloofwaardige vals een te onderskei self ’n moeilike probleem is, en dat om *net* ware stellings te lewer minstens net so moeilik is. ’n Vloer van foute is ingebou.

Die idee daaronder: hoe om te tel wat jy nog nie gesien het nie

Dit rus op ’n werklik slim idee — ouer as taalmodelle, en die moeite werd om behoorlik te ontmoet.

Begin met ’n sak gekleurde balle. Jy weet nie hoeveel kleure dit bevat nie. Jy trek 100, een op ’n slag, en hou tellings: rooi 40, blou 25, groen 15, geel 5, pers 3, oranje 2 — en dan **tien verskillende kleure wat elk presies een keer opduik.**

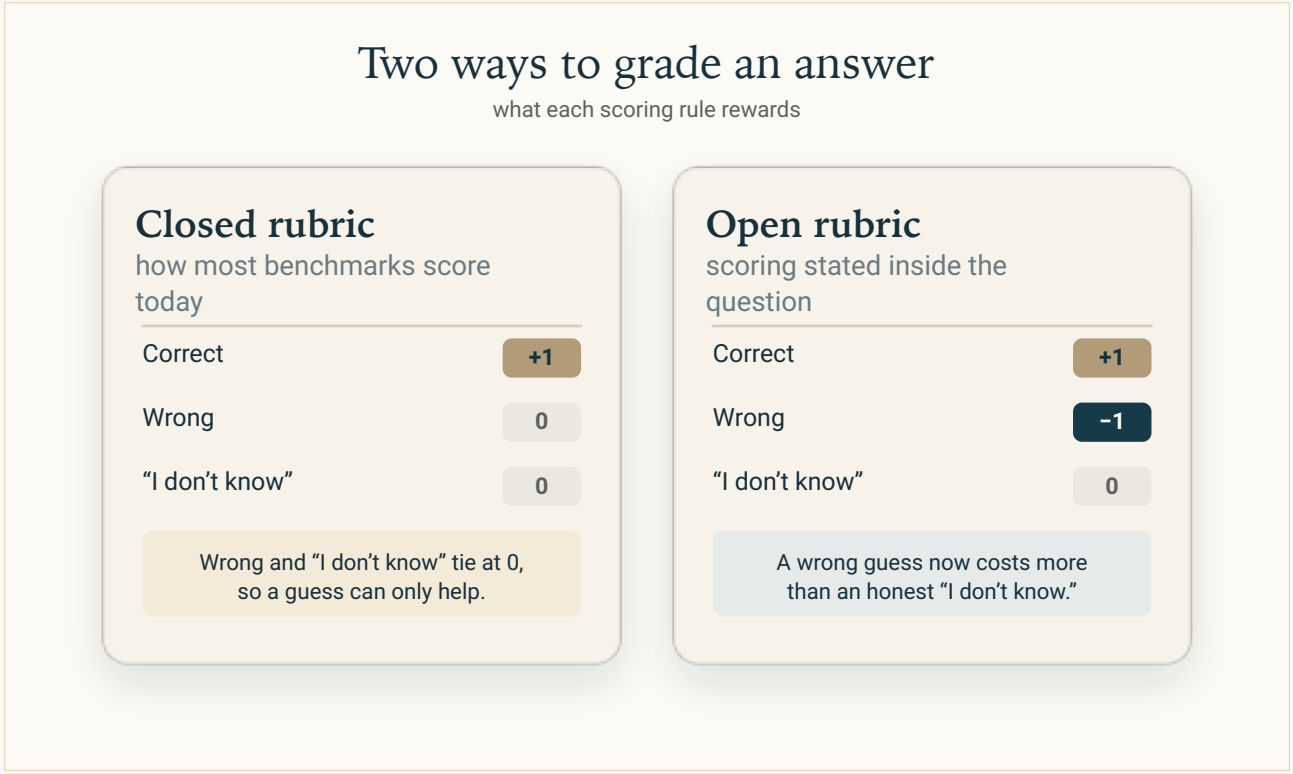
Nou die vraag waarvoor Turing werklik te staan gekom het, by ’n heel ander probleem: *wat is die kans dat die volgende bal ’n kleur is wat jy glad nie gesien het nie?* Jy kan nie tel wat jy nooit getrek het nie — maar jy **kan** die kleure tel wat jy presies een keer gesien het, die “enkelinge”. Die slenter, **Good-Turing-skatting** genoem, is dat die aandeel van jou trekke wat enkelinge is, die waarskynlikheid skat wat nog wegkruip in die kleure wat jy nie gesien het nie. Tien van jou honderd trekke was eenmalige kleure, so die kans dat die volgende bal ’n splinternuwe kleur is, is omtrent **10 / 100 = 10%.**

Daardie een keer gesiene kleure is nie *foute* nie. Hulle is ’n meting van jou eie onkunde: baie kleure wat een keer opduik, is die steekproef se manier om jou te vertel dat die wêreld meer bevat as wat jy eenvoudig nog nie getrek het nie.

Ruil nou kleure vir verjaardae, en die sak vir die model se opleidingsteks. Gestel, onder die verjaardae wat dit gesien het, verskyn **een uit vyf presies een keer.** Dieselfde slenter: omtrent ’n vyfde van die waarskynlikheid lê in verjaardae wat die model effektief nooit gesien het nie — en ’n verjaardag het geen patroon om op terug te val nie (jy kan nie iemand se verjaardag *uitwerk* nie). ’n Datum wat een keer of nooit gesien is, is dus ’n muntstuk wat die model nie kan

> weeg nie, en dit sal by ruweg **een uit vyf** daarvan verkeerd wees. Geen hoeveelheid slimheid maak dit reg nie: daar was niks om te leer nie.
> Dit is die hele argument in miniatuur: die enkeling-koers meet hoeveel van die wêreld *uit hierdie data* onleerbaar is, en dit word 'n **vloer** onder die foute. Dit is ook waarom 'n model byna nooit 'n hoofstad mis nie — *Parys* verskyn aanhoudend, sy enkeling-koers is naby nul, so daar is oorgenoeg om te leer.

Dit verklaar waar hallusinasies vandaan kom. Dit verklaar nie waarom hulle *oorleef* nie — waarom modelle, ná al die latere opleiding wat hulle behulpsaam en eerlik moet maak, steeds bluf eerder as om twyfel te erken. Hier is die artikel se analogie amper ongemaklik raak. Stel jou 'n student in 'n eksamen voor wat 'n antwoord nie ken nie. As 'n leë blok nul punte kry en 'n raaiskoot dalk een, is die puntemaksimerende skuif om te raai — met selfvertroue, spesifiek, nooit "ek is nie seker nie". Studente leer dit. En modelle, blyk dit, ook — want ons gee hulle punte op dieselfde manier. Die outeurs het die maatstawwe deurgegaan waarop die veld werklik meeding, die ranglyste waarvoor modelle ingestel word, en gevind dat byna almal "ek weet nie" presies dieselfde telling gee as 'n verkeerde antwoord: nul. Onder daardie reël sal 'n model wat altyd raai 'n andersins identiese model klop wat sy onsekerheid eerlik aandui. Ons gee hulle, taamlik letterlik, punte daarvoor in.



Maatstawwe kan raai rasioneel maak. Onder die puntetoekenning wat die meeste maatstawwe gebruik (links), kry 'n verkeerde antwoord en 'n eerlike "ek weet nie" albei nul — 'n raaiskoot kan dus net help. Noem die reëls in die vraag self, met 'n strafpunt vir verkeerd (regs), en om jou by onsekerheid te weerhou kan die beter skuif word. Dit verander wat die toets beloon; dit los hallusinasie nie op sy eie op nie.

Dit is die deel wat 'n mens moet hou, want dit loop teen die gewone opskrif in. Hallusinasie word

dikwels verkoop as ’n onvermydelike, byna mistieke grens van die tegnologie. Die artikel betwis albei. Die vooropleidingsvloer is geen misterie nie — dit is gewone statistiese fout, van die soort wat masjienleer al dekades lank verstaan. En die voortbestaan is nie onvermydelik nie — dit is, deels, ’n aansporing wat ons gebou het en kan verander. ’n Stelsel wat by onsekerheid eenvoudig weier om te antwoord, sou glad nie hallusineer nie; die rede waarom uitgerolde modelle nie so optree nie, is dat ons telborde die weiering straf.

Wat die outeurs gedoen het

Die artikel het drie dele. Eerstens ’n wiskundige argument dat ’n deel van hallusinasie statisties afgedwing word tydens vooropleiding, deur te wys dat “genereer net geldige teks” minstens so moeilik is soos ’n binêre klassifikasieprobleem “is hierdie stelling geldig?”. Tweedens ’n argument — gestaaf deur ’n oorsig van tien invloedryke maatstawwe — dat hoofstroom-akkuraatheidsmetrieke raai bo weerhouding beloon. Derdens ’n voorgestelde oplossing en ’n gevallestudie wat dit toets: **oop rubrieke**, waar die puntetoekenning binne die vraag self genoem word (byvoorbeeld: “’n regte antwoord tel 1, ’n verkeerde −1, weerhou jou dus as jy minder as 50% seker is”), sodat ’n model kan agterkom wanneer eerlikheid beloon word. Hulle probeer dit op vier voorste modelle — Google se Gemini 3 Pro, OpenAI se GPT-5, xAI se Grok 4 en Anthropic se Claude Opus 4.5 — met SimpleQA se 4 326 feitevrae. Hulle is uitdruklik daaroor dat die gevallestudie illustratief is, “nie ’n gekontroleerde evaluasie oor modelle heen nie” (verstekinstellings, geen instelling, geen kostenormalisering nie).

Wat hulle gevind het

- **Vooropleiding dwing ’n deel van die foute af.** Die tempo waarteen ’n model selfversekerde onwaarhede lewer, word van onder begrens deur (ruweg twee keer) die foutkoers van die beste “is hierdie stelling geldig?”-klassifiseerder wat daaruit gebou kan word. Vir feite sonder ’n leerbare patroon lê daardie vloer minstens by die *enkeling-koers* — die breukdeel feite wat presies een keer in die opleiding voorkom. ’n Deel van hallusinasie is onvermydelik, selfs met volmaak skoon data.
- **Puntetoekenning beloon raai — konkreet.** Onder gewone reg/verkeerd-telling is nooit-weerhou die optimale strategie, en die outeurs se oorsig vind dat die oorgrote meerderheid gewilde maatstawwe “ek weet nie” bloot as verkeerd tel. ’n Treffende voorbeeld uit eie geledere: op die SimpleQA-toets *bevoordeel* rou akkuraatheid OpenAI se o4-mini effens — wat byna alles beantwoord en meer as driekwart van die tyd verkeerd is — bo GPT-5-mini, wat veel minder foute maak omdat dit hom by onsekerheid weerhou. Die roekelose model lyk beter op die telbord.
- **Oop rubrieke keer die aansporing om (in hulle gevallestudie).** Hulle toets ’n eenvoudige hallusinasie-verligting (laat die model twee keer antwoord en weerhou as die twee antwoorde verskil). Onder standaardakkuraatheid sny die verligting foute *maar ook akkuraatheid* — die metriek ontmoedig dus die aanvaarding daarvan. Onder oop rubrieke kom dieselfde verligting by al vier modelle voor oor ’n reeks strafwaardes; en GPT-5-mini — wat rou akkuraatheid gestraf het omdat dit hom by onsekerheid weerhou — kom bo o4-mini uit sodra die puntetoekenning openlik genoem word (n = 4 326 vrae per model).

Wat dit waarskynlik beteken

Om hallusinasie te verminder is meestal nie ’n kwessie van nog hallusinasie-spesifieke toetse uitdink nie. Dit is ’n kwessie van verander hoe die hoofstroommaatstawwe onsekerheid tel, sodat om “ek weet nie” te erken nie meer gestraf word nie. Totdat die telbord verander, sal die vermindering van hallusinasie modelle akkuraatheidspunte bly kos en dus ontmoedig bly — en daarom raam die outeurs die probleem as “sosio-tegnies”: deels ’n beter metriek, deels om die invloedryke ranglyste sover te kry om dit aan te neem.

Wat dit *nie* bewys nie

- Dit wys **nie** dat oop rubrieke hallusinasie in die praktyk regmaak nie. Die ondersteunende eksperiment is ’n klein, doelbewus **ongekontroleerde** gevallestudie — vier modelle op verstekinstellings, een gekose verligting, een feite-toets — bedoel om die *aansporing-omkeer* te demonstreer, nie om modelle te rangskik of algemene doeltreffendheid te bewys nie.
- Dit beweer **nie** dat puntetoekenning die *enigste* oorsaak is nie. Foute in die opleidingsdata, werklik moeilike probleme en onbekende opdragte bly aparte bronne.
- Dit ondersteun **nie** die gewilde lyn dat hallusinasies onvermydelik is nie. Die outeurs voer die teenoorgestelde aan: ’n stelsel wat net nagaanbare vrae beantwoord en andersins “ek weet nie” sê, sou nooit hallusineer nie.
- Dit laat die vooropleidingsvloer **nie** verdwyn nie — dit verklaar en begrens dit, en die grens gaan oor selfversekerde feitefoute, nie oor alle modelgedrag nie.
- Dit wys **nie** dat oop rubrieke op hulle eie *voldoende* is nie. Hulle verander wat ’n evaluasie beloon; hulle is geen plaasvervanger vir herwinning, gereedskapsgebruik of beter gekalibreerde modelle nie.

Hoe sterk is die getuienis?

- Die kern is **wiskunde** — formele ondergrense, nie metings nie. As teoretiese argument is dit op sy eie terme deeglik.
- Dit rus op **doelbewus vereenvoudigde modelle** van die probleem; die outeurs self merk die “vals trigotomie” aan om elke antwoord as reg, verkeerd of “ek weet nie” te behandel, en die geïdealiseerde opset van “arbitrêre feite” vir die skoonste grens.
- Die maatstaf-oorsig is ’n **klein, gekureerde steekproef** — tien invloedryke evaluasies, nie ’n uitputtende oudit nie.
- Die gevallestudie is **eg maar beperk**: vier voorste modelle, ’n enkele verligting, net SimpleQA, verstekinstellings, uitdruklik “nie ’n gekontroleerde evaluasie nie”. Dit is ’n konsepbewys vir die aansporingsargument, nie ’n maatstafresultaat nie.
- Die uitkykpunt verdien om genoem te word: drie van die vier outeurs werk of het gewerk by **OpenAI**, en die artikel voer aan dat die veld moet verander hoe dit modelle evalueer. Dit is ’n

goed beredeneerde standpunt van ’n belanghebbende party, nie ’n neutrale buiteblik nie — om te weeg, nie om af te wys nie. (Tot sy krediet rig die artikel die kritiek net so geredelik op sy eie modelle, o4-mini en GPT-5-mini, as op ander.)

Waarom dit saak maak

Dit herraam ’n swaar opgeblaasde probleem. “Hallusinasie” word gewoonlik verkoop as óf ’n spookagtige defek óf ’n onwrikbare muur; hierdie artikel maak dit gewoon en deels selfopgelê — ’n statistiese vloer wat ons werklik kan verstaan, bo-op ’n aansporing wat ons gekies het. Die wyer les is stiller en nuttiger: verdere vordering met betroubaarheid kan net soveel afhang van *wat ons meet* as van wat ons bou.

Skoon opsomming

Selfversekerde vals antwoorde van taalmodelle kom uit twee plekke. Die eerste is statisties: wanneer ’n feit geen leerbare patroon het nie, sal ’n model wat taal moet naboots dit soms verkeerd kry, en daardie vloer kan geskat word (byvoorbeeld uit hoeveel feite net een keer in die opleiding voorkom). Die tweede is aansporings: byna elke maatstaf waarop modelle gerangskik word, tel “ek weet nie” dieselfde as ’n verkeerde antwoord, so raai wen altyd — tot op die punt dat ’n model wat driekwart van die tyd verkeerd is, ’n eerliker een kan uitstof wat hom weerhou. Die outeurs se voorstel is nie nóg ’n hallusinasietoets nie, maar “oop rubrieke”: noem die puntetoekenning binne die vraag. In ’n gevallestudie op vier voorste modelle keer dit die aansporing om, sodat ’n hallusinasie-verminderende metode beloon eerder as gestraf word. Dit is ’n teorie-plus-oorsig-artikel met ’n klein, uitdruklik ongekontroleerde eksperiment, peer-reviewed in *Nature*; die oplossing is belowend maar nog nie op skaal bewys nie, en hallusinasies is, so lui die betoog, nóg misterieus nóg streng onvermydelik.

No-BS-toets

Wat die artikel wys: ’n Wiskundige ondergrens waaronder ’n deel van hallusinasie tydens vooropleiding afgedwing word (minstens die “enkeling-koers” by patroonlose feite); ’n oorsig wat vind dat die meeste voorste maatstawwe “ek weet nie” geen krediet gee nie; en ’n gevallestudie met vier modelle waarin die noem van die puntetoekenning in die opdrag (“oop rubrieke”) ’n hallusinasie-verminderende metode laat wen waar rou akkuraatheid dit gestraf het.

Wat aanneemlik maar nie bewys is nie: Dat oop rubrieke, in hoofstroommaatstawwe opgeneem, hallusinasie in uitgerolde modelle merkbaar sou verminder. Die ondersteunende eksperiment is klein en uitdruklik ongekontroleer.

Wat dit nie wys nie: Dat hallusinasies onvermydelik is (dit voer die teenoorgestelde aan); dat puntetoekenning die enigste oorsaak is; dat hallusinasie heeltemal uitgeskakel kan word; dat die gevallestudie die vier modelle teen mekaar rangskik.

Vernaamste beperkings: ’n Doelbewus vereenvoudigde reg/verkeerd/“ek weet nie”-model (die outeurs noem dit ’n “vals trigotomie”); ’n klein gekureerde maatstaf-oorsig (tien evaluasies); ’n ongekontroleerde gevallestudie (vier modelle, een verligting, een toets, verstekinstellings); en ’n OpenAI-geleide betoog oor hoe die veld modelle behoort te evalueer.

Hoeveel vertroue behoort ’n algemene leser te hê? Baie daarin dat hallusinasie nóg misterieus nóg streng onvermydelik is, en dat hoofstroommaatstawwe tans raai beloon. Matig daarin dat die voorgestelde oplossing help — daar is nou ’n egte konsepbewys, maar nog geen demonstrasie dat dit breed en op skaal werk nie.

Sources

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