The Clean Paper

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

Waarom taalmodellen hallucineren — en waarom onze beoordeling het zo houdt

De stellige onwaarheden die we "hallucinaties" noemen zijn geen mysterieuze storing: een deel is een statistisch bijproduct van de training, en ze blijven bestaan omdat gangbare benchmarks een zelfverzekerde gok hoger belonen dan een eerlijk "ik weet het niet". Een casestudy op vier frontier-modellen laat zien dat het vermelden van de scoreregels in de prompt ("open rubrics") die prikkel omkeert.

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

Waarom modellen gokken — en waarom wij ze dat hebben aangeleerd

Vraag een groot taalmodel naar de verjaardag van een onbekende en het antwoordt misschien “7 maart” — met de kalme stelligheid van iemand die het van een kaartje afleest — en zit ernaast, drie keer op rij, met drie verschillende datums. De auteurs geven precies dit soort voorbeelden: toon-aangevende modellen die een gewone feitenvraag kregen — iemands verjaardag, of waar een obscuur acroniem voor staat — en elk vol zelfvertrouwen een ander antwoord verzonnen, geen ervan juist. Het woord van de industrie hiervoor is *hallucinatie*, wat klinkt als een waarnemingsstoornis. De eerste zet van de paper is het mysterie eruit halen.

Begin bij hoe een model wordt gebouwd. In zijn eerste en grootste trainingsfase leert het, in feite, hoe vloeiende taal eruitziet door een enorme hoeveelheid tekst te lezen. Neem nu een feit zonder patroon erachter — de verjaardag van één bepaalde persoon. Als die datum één keer in de trainingstekst voorkwam, of nooit, dan is er voor een patroonleerder niets om vast te pakken: het antwoord is, vanuit het model gezien, willekeurig. De auteurs maken dit precies door een oud idee te lenen (van Alan Turing, voor een ander probleem): als één op de vijf verjaardagen maar één keer in de data opduikt, mag je verwachten dat een model er minstens één op de vijf fout heeft — niet omdat het kapot is, maar omdat er nooit iets te leren viel. (Volgens dezelfde logica zit een model er bij de hoofdstad van een land vrijwel nooit naast: die komen voortdurend voor.) Ze betogen, met de nodige zorg, dat een ware bewering onderscheiden van een plausibele onware er zelf al een hard probleem is, en dat *alleen* ware beweringen produceren minstens zo moeilijk is. Een bodem aan fouten is ingebakken.

Het idee eronder: tellen wat je nog niet hebt gezien

Dit rust op een werkelijk slim idee — ouder dan taalmodellen, en het verdient een echte kennismaking.

Begin met een zak gekleurde ballen. Je weet niet hoeveel kleuren erin zitten. Je trekt er 100, één voor één, en turft: rood 40, blauw 25, groen 15, geel 5, paars 3, oranje 2 — en dan **tien verschillende kleuren die elk precies één keer opduiken.**

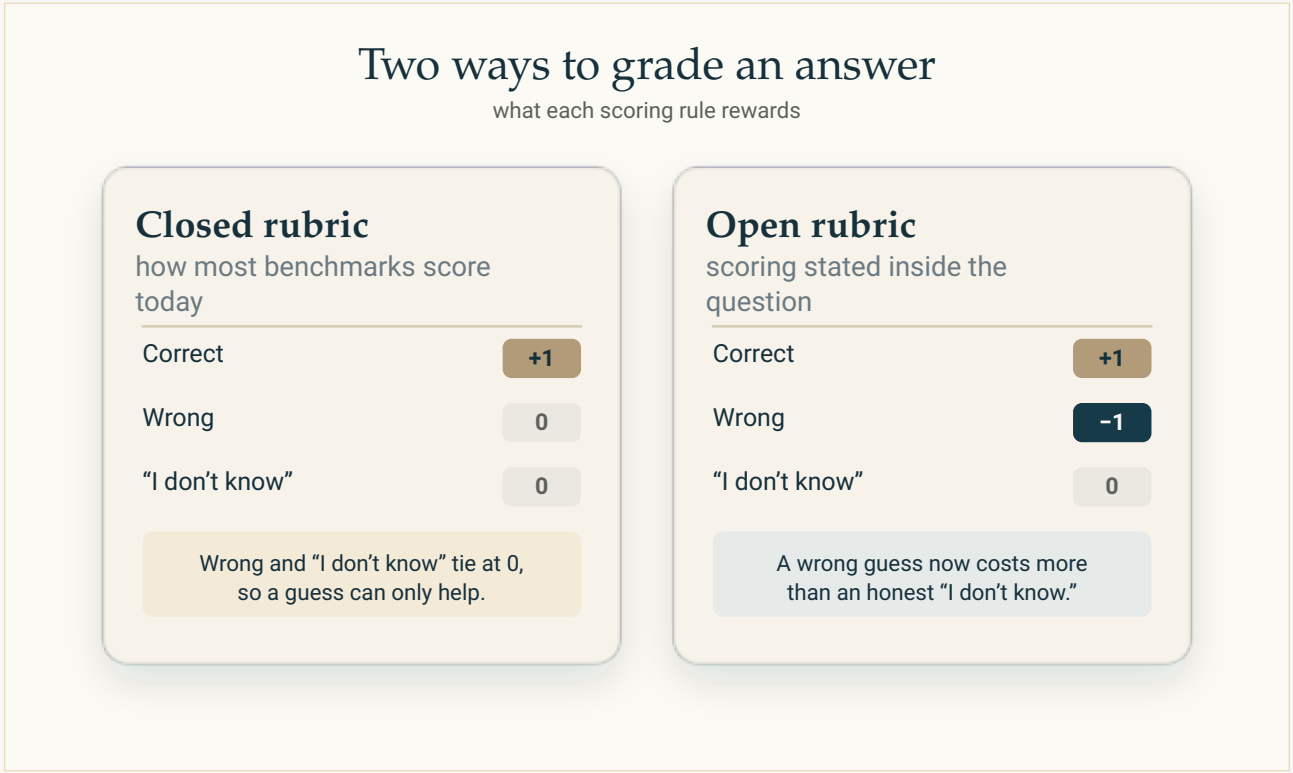
Nu de vraag waar Turing werkelijk voor stond, bij een heel ander probleem: *hoe groot is de kans dat de volgende bal een kleur heeft die je nog helemaal niet hebt gezien?* Je kunt niet tellen wat je nooit hebt getrokken — maar je **kunt** de kleuren tellen die je precies één keer hebt gezien, de “singletons”. De truc, **Good-Turing-schatting** genoemd, is dat het aandeel van je trekkingen dat singleton is, de kans schat die nog verborgen zit in de kleuren die je niet hebt gezien. Tien van je honderd trekkingen waren eenmalige kleuren, dus de kans dat de volgende bal een gloednieuwe kleur heeft is ongeveer **10 / 100 = 10%.**

Die één keer geziene kleuren zijn geen *fouten*. Ze zijn een meting van je eigen onwetendheid: veel kleuren die één keer opduiken is de manier waarop de steekproef je vertelt dat de wereld meer bevat dan je simpelweg nog niet hebt getrokken.

Verwissel nu kleuren voor verjaardagen, en de zak voor de trainingstekst van het model. Stel dat, onder de verjaardagen die het zag, **één op de vijf precies één keer voorkomt.** Zelfde truc: ongeveer een vijfde van de kans zit in verjaardagen die het model feitelijk nooit heeft gezien — en een verjaardag heeft geen patroon om op terug te vallen (je kunt iemands verjaardag niet

> *beredeneren*). Een datum die één keer of nooit is gezien is dus een munt die het model niet kan wegen, en het zal er bij ruwweg **één op de vijf** naast zitten. Geen enkele slimheid lost dit op: er viel niets te leren.
> Dat is het hele argument in miniatuur: het singleton-aandeel meet hoeveel van de wereld *uit deze data* onleerbaar is, en dat wordt een **bodem** onder de fouten. Het is ook waarom een model vrijwel nooit een hoofdstad mist — *Parijs* komt voortdurend voor, zijn singleton-aandeel is bijna nul, dus er valt volop te leren.

Dat verklaart waar hallucinaties vandaan komen. Het verklaart niet waarom ze *overleven* — waarom modellen, na al de latere training die ze behulpzaam en eerlijk moet maken, nog steeds bluffen in plaats van twijfel toe te geven. Hier is de analogie van de paper bijna ongemakkelijk raak. Stel je een student voor in een tentamen die een antwoord niet weet. Als een leeg vak nul punten oplevert en een gok er misschien één, dan is de cijfermaximaliserende zet: gokken — zelfverzekerd, specifiek, nooit “ik weet het niet zeker”. Studenten leren dat. En modellen, zo blijkt, ook — omdat wij ze op dezelfde manier beoordelen. De auteurs liepen de benchmarks na waarop het veld daadwerkelijk concurreert, de ranglijsten waarvoor modellen worden bijgesteld, en vonden dat vrijwel allemaal ze “ik weet het niet” exact dezelfde score geven als een fout antwoord: nul. Onder die regel verslaat een model dat altijd gokt een verder identiek model dat zijn onzekerheid eerlijk aangeeft. We scoren ze er, vrij letterlijk, in.



Benchmarks kunnen gokken rationeel maken. Onder de score die de meeste benchmarks gebruiken (links) leveren een fout antwoord en een eerlijk “ik weet het niet” allebei nul op — een gok kan dus alleen maar helpen. Zet de regels in de vraag zelf, met een straf voor fout (rechts), en je onthouden bij onzekerheid kan de betere zet worden. Het verandert wat de test beloont; het lost hallucinatie op zichzelf niet op.

Dit is het deel om te onthouden, want het gaat in tegen de gebruikelijke kop. Hallucinatie wordt vaak verkocht als een onvermijdelijke, bijna mystieke grens van de technologie. De paper bestrijdt beide. De pretraining-bodem is geen mysterie — het is gewone statistische fout, van het soort dat machine learning al decennia begrijpt. En het voortduren is niet onvermijdelijk — het is, deels, een prikkel die wij hebben gebouwd en kunnen veranderen. Een systeem dat bij onzekerheid simpelweg weigerde te antwoorden zou helemaal niet hallucineren; de reden dat uitgerolde modellen zich niet zo gedragen is dat onze scoreborden de weigering afstraffen.

Wat de auteurs deden

De paper heeft drie delen. Ten eerste een wiskundig argument dat een deel van de hallucinatie statistisch wordt afgedwongen tijdens pretraining, door te laten zien dat “genereer alleen geldige tekst” minstens zo moeilijk is als een binair classificatieprobleem “is deze bewering geldig?”. Ten tweede een argument — gestaafd door een doorlichting van tien invloedrijke benchmarks — dat gangbare nauwkeurigheidsmetrieken gokken belonen boven je onthouden. Ten derde een voorgestelde oplossing plus een casestudy die haar test: **open rubrics**, waarbij de score in de vraag zelf staat (bijvoorbeeld: “een juist antwoord scoort 1, een fout −1, onthoud je dus als je minder dan 50% zeker bent”), zodat een model kan zien wanneer eerlijkheid wordt beloond. Ze proberen dit op vier frontiermodellen — Googles Gemini 3 Pro, OpenAI’s GPT-5, xAI’s Grok 4 en Anthropics Claude Opus 4.5 — met de 4.326 feitenvragen van SimpleQA. Ze zijn er expliciet over dat de casestudy illustratief is, “geen gecontroleerde evaluatie over modellen heen” (standaardinstellingen, geen tuning, geen kostennormalisatie).

Wat ze vonden

- **Pretraining dwingt een deel van de fouten af.** Het tempo waarin een model stellige onwaarheden produceert is van onderen begrensd door (ruwweg twee keer) de foutkans van de beste “is deze bewering geldig?”-classificator die eruit te bouwen valt. Voor feiten zonder leerbaar patroon ligt die bodem minstens op het *singleton-aandeel* — de fractie feiten die precies één keer in de training voorkomt. Een deel van de hallucinatie is onvermijdelijk, zelfs met perfect schone data.
- **Beoordeling beloont gokken — concreet.** Onder gewone goed/fout-scoring is nooit-onthouden de optimale strategie, en de doorlichting van de auteurs vindt dat de overgrote meerderheid van populaire benchmarks “ik weet het niet” gewoon als fout scoort. Een sprekend voorbeeld uit eigen huis: op de SimpleQA-test *bevoordeelt* de kale nauwkeurigheid OpenAI’s o4-mini licht — dat vrijwel alles beantwoordt en er in meer dan driekwart van de gevallen naast zit — boven GPT-5-mini, dat veel minder fouten maakt omdat het zich bij onzekerheid onthoudt. Het roekelozere model oogt beter op het scorebord.
- **Open rubrics keren de prikkel om (in hun casestudy).** Ze testen een eenvoudige hallucinatiemitigatie (laat het model twee keer antwoorden en onthoud je als de twee antwoorden verschillen).

Onder standaardnauwkeurigheid snijdt de mitigatie fouten weg *maar ook nauwkeurigheid* — de metriek ontmoedigt dus haar invoering. Onder open rubrics komt dezelfde mitigatie bij alle vier de modellen als winnaar uit de bus over een reeks strafwaarden; en GPT-5-mini — dat door kale nauwkeurigheid was afgestraft voor het zich onthouden bij twijfel — eindigt boven o4-mini zodra de score open wordt vermeld (n = 4.326 vragen per model).

Wat dit waarschijnlijk betekent

Hallucinatie terugdringen is vooral geen kwestie van meer hallucinatie-specifieke tests verzinnen. Het is een kwestie van veranderen hoe de gangbare benchmarks onzekerheid scoren, zodat "ik weet het niet" toegeven niet langer wordt bestraft. Tot het scorebord verandert, zal hallucinatie verminderen modellen nauwkeurigheidspunten blijven kosten en dus ontmoedigd blijven — en daarom framen de auteurs het probleem als "sociotechnisch": deels een betere metriek, deels de invloedrijke ranglijsten zover krijgen die over te nemen.

Wat dit *niet* bewijst

- Het toont **niet** aan dat open rubrics hallucinatie in het wild verhelpen. Het ondersteunende experiment is een kleine, bewust **ongecontroleerde** casestudy — vier modellen op standaardinstellingen, één gekozen mitigatie, één feiten-QA-test — bedoeld om de *prikkelomkering* te demonstreren, niet om modellen te rangschikken of algemene werkzaamheid te bewijzen.
- Het claimt **niet** dat beoordeling de *enige* oorzaak is. Fouten in de trainingsdata, werkelijk moeilijke problemen en onbekende prompts blijven aparte bronnen.
- Het steunt **niet** de populaire lijn dat hallucinaties onvermijdelijk zijn. De auteurs betogen het tegendeel: een systeem dat alleen controleerbare vragen beantwoordde en anders "ik weet het niet" zei, zou nooit hallucineren.
- Het laat de pretraining-bodem **niet** verdwijnen — het verklaart en begrenst hem, en de grens betreft stellige feitelijke fouten, niet al het modelgedrag.
- Het toont **niet** aan dat open rubrics op zichzelf *volstaan*. Ze veranderen wat een evaluatie beloont; ze zijn geen vervanging voor retrieval, gereedschapsgebruik of beter gekalibreerde modellen.

Hoe sterk is het bewijs?

- De kern is **wiskunde** — formele ondergrenzen, geen metingen. Als theoretisch argument is het op eigen termen deugdelijk.
- Het rust op **bewust vereenvoudigde modellen** van het probleem; de auteurs benoemen zelf de "valse trichotomie" van elke respons behandelen als goed, fout of "ik weet het niet", en de geïdealiseerde setting van "willekeurige feiten" voor de schoonste grens.
- De benchmark-doorlichting is een **kleine, samengestelde steekproef** — tien invloedrijke evalu-

aties, geen uitputtende audit.
- De casestudy is **echt maar beperkt**: vier frontier-modellen, één enkele mitigatie, alleen SimpleQA, standaardinstellingen, expliciet "geen gecontroleerde evaluatie". Het is een proof of concept voor het prikkelargument, geen benchmarkresultaat.
- Het gezichtspunt verdient benoeming: drie van de vier auteurs werken of werkten bij **OpenAI**, en de paper betoogt dat het veld moet veranderen hoe het modellen evalueert. Dat is een goed beargumenteerde positie van een belanghebbende partij, geen neutrale blik van buiten — om te wegen, niet om weg te wuiven. (Sier het gezegd: de paper richt de kritiek even gemakkelijk op de eigen modellen, o4-mini en GPT-5-mini, als op andere.)

Waarom het ertoe doet

Het herkadert een zwaar gehyped probleem. "Hallucinatie" wordt doorgaans verkocht als een griezelig defect of als een onwrikbare muur; deze paper maakt haar gewoon en deels zelf toegebracht — een statistische bodem die we werkelijk kunnen begrijpen, bovenop een prikkel die we zelf hebben gekozen. De bredere les is stiller en nuttiger: verdere vooruitgang in betrouwbaarheid kan evenzeer afhangen van *wat we meten* als van wat we bouwen.

Heldere samenvatting

Stellige onjuiste antwoorden van taalmodellen komen uit twee bronnen. De eerste is statistisch: wanneer een feit geen leerbaar patroon heeft, zal een model dat taal moet nabootsen het soms fout hebben, en die bodem is te schatten (bijvoorbeeld aan de hand van hoeveel feiten maar één keer in de training voorkomen). De tweede zijn prikkels: vrijwel elke benchmark waarop modellen worden gerangschikt scoort "ik weet het niet" hetzelfde als een fout antwoord, dus gokken wint altijd — tot het punt dat een model dat er driekwart van de tijd naast zit een eerlijker model kan overtreffen dat zich onthoudt. Het voorstel van de auteurs is geen zoveelste hallucinatietest maar "open rubrics": zet de score in de vraag zelf. In een casestudy op vier frontier-modellen keert dat de prikkel om, zodat een hallucinatie-verminderende methode wordt beloond in plaats van bestraft. Het is een theorie-plus-doorlichting-paper met een klein, expliciet ongecontroleerd experiment, peer-reviewed verschenen in *Nature*; de oplossing is veelbelovend maar nog niet op schaal aangetoond, en hallucinaties zijn, zo luidt het betoog, mysterieus noch strikt onvermijdelijk.

Zonder omwegen

Wat de paper aantoont: Een wiskundige ondergrens waaronder een deel van de hallucinatie tijdens pretraining wordt afgedwongen (minstens het "singleton-aandeel" bij patroonloze feiten); een doorlichting die vindt dat de meeste toonaangevende benchmarks "ik weet het niet" geen enkel punt geven; en een casestudy met vier modellen waarin het vermelden van de score in de prompt ("open

rubrics”) een hallucinatie-verminderende methode laat winnen waar kale nauwkeurigheid haar had bestraft.

Wat plausibel maar niet bewezen is: Dat open rubrics, opgenomen in gangbare benchmarks, hallucinatie in uitgerolde modellen merkbaar zouden verminderen. Het ondersteunende experiment is klein en expliciet ongecontroleerd.

Wat het niet aantoont: Dat hallucinaties onvermijdelijk zijn (het betoogt het omgekeerde); dat beoordeling de enige oorzaak is; dat hallucinatie volledig valt uit te bannen; dat de casestudy de vier modellen onderling rangschikt.

Belangrijkste beperkingen: Een bewust vereenvoudigd goed/fout/“ik weet het niet”-model (de auteurs noemen het een “valse trichotomie”); een kleine samengestelde benchmark-doorlichting (tien evaluaties); een ongecontroleerde casestudy (vier modellen, één mitigatie, één test, standaardinstellingen); en een door OpenAI geleid betoog over hoe het veld modellen zou moeten evalueren.

Hoeveel vertrouwen mag een algemene lezer hebben? Veel in dat hallucinatie mysterieus noch strikt onvermijdelijk is, en dat gangbare benchmarks gokken momenteel belonen. Gematigd in dat de voorgestelde oplossing helpt — er ligt nu een echt proof of concept, maar nog geen demonstratie dat het breed en op schaal werkt.

Bronnen

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