The Clean Paper

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

De ce halucinează modelele lingvistice și de ce felul în care le evaluăm menține problema

Falsurile rostite cu siguranță, pe care le numim "halucinații", nu sunt o eroare misterioasă: unele sunt un produs secundar statistic al antrenării și persistă pentru că benchmarkurile dominante recompensează o presupunere sigură pe sine mai mult decât un onest "nu știu". Un studiu de caz pe patru modele frontier arată că precizarea regulilor de scor în prompt ("rubrici deschise") inversează acest stimulent.

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

De ce modelele ghicesc și de ce le-am învățat noi să facă asta

Întreabă un model lingvistic de mari dimensiuni care este ziua de naștere a unei persoane necunoscute și s-ar putea să răspundă „7 martie” cu siguranța calmă a cuiva care citește de pe o fișă — și să greșească, de trei ori la rând, cu trei date diferite. Autorii dau exact acest tip de exemplu: modele de vârf primesc o întrebare factuală simplă — ziua de naștere a unei persoane sau semnificația unui acronim obscur — și fiecare inventează cu încredere un răspuns diferit, niciunul corect. Cuvântul industriei pentru asta este *halucinație*, ceea ce face fenomenul să sune ca o defecțiune de percepție. Prima mișcare a lucrării este să-i ia misterul.

Să începem cu felul în care este construit un model. În prima și cea mai mare etapă de antrenare, el învață, în esență, cum arată limbajul fluent citind o cantitate enormă de text. Acum ia un fapt care nu are un tipar în spate — ziua de naștere a unei persoane anume. Dacă data respectivă a apărut o singură dată în textul de antrenare, sau nu a apărut deloc, nu există nimic de care un sistem care învață tipare să se poată agăța: din punctul de vedere al modelului, răspunsul este arbitrar. Autorii fac această idee precisă împrumutând un argument vechi (al lui Alan Turing, dintr-o problemă diferită): dacă una din cinci zile de naștere apare o singură dată în date, ar trebui să ne așteptăm ca un model să greșească cel puțin una din cinci — nu pentru că este stricat, ci pentru că nu era nimic acolo de învățat. (După aceeași logică, modelele aproape nu greșesc niciodată capitala unei țări: acelea apar continuu.) Autorii susțin, cu grijă, că a deosebi o afirmație adevărată de una falsă dar plauzibilă este deja o problemă dificilă, iar a produce *numai* afirmații adevărate este cel puțin la fel de dificil. Există un prag minim de eroare construit în sistem.

Ideea de dedesubt: cum numeri ce nu ai văzut încă

Aici stă o idee cu adevărat inteligentă — mai veche decât modelele lingvistice, și merită întâlnită cum trebuie.

Începe cu un sac de bile colorate. Nu știi câte culori conține. Scoți 100, una câte una, și le numeri: roșii 40, albastre 25, verzi 15, galbene 5, mov 3, portocalii 2 — și apoi **zece culori diferite care apar exact o dată fiecare.**

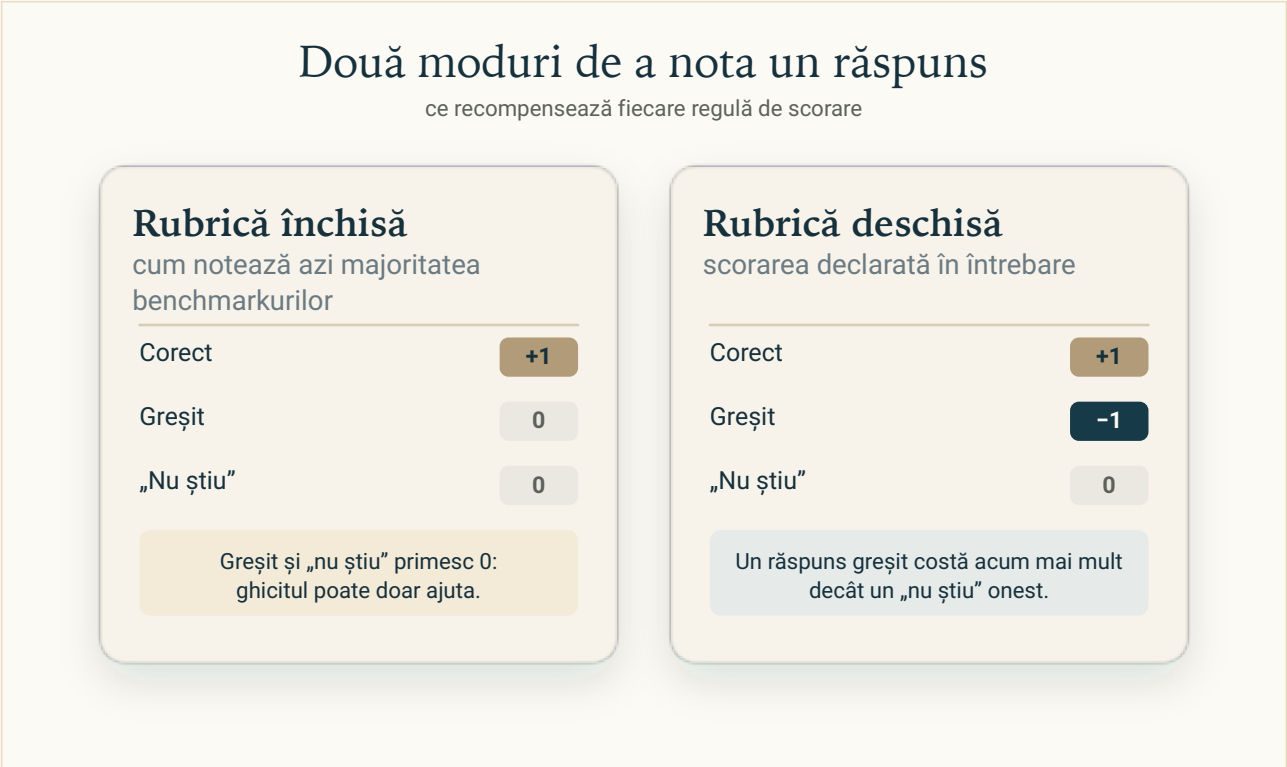
Acum întrebarea pe care Turing chiar a avut-o în față, într-o problemă cu totul diferită: *care este șansa ca următoarea bilă să fie de o culoare pe care nu ai văzut-o deloc?* Nu poți număra ce nu ai extras niciodată — dar poți număra culorile pe care le-ai văzut exact o dată, „singletonurile”. Trucul, numit **estimarea Good–Turing**, este că ponderea extragerilor tale care sunt singletonuri estimează probabilitatea încă ascunsă în culorile pe care nu le-ai văzut. Zece dintre cele o sută de extrageri au fost culori apărute o singură dată, deci șansa ca următoarea bilă să fie o culoare complet nouă este cam **10 / 100 = 10%.**

Acele culori văzute o singură dată nu sunt *greșeli*. Sunt o măsură a propriei tale ignoranțe: multe culori care apar o dată sunt felul în care eșantionul îți spune că lumea conține mai mult decât ai extras până acum.

Acum înlocuiește culorile cu zile de naștere, iar sacul cu textul de antrenare al modelului. Să presupunem că, dintre zilele de naștere pe care le-a văzut, **una din cinci apare exact o dată.** Același truc: cam o cincime din probabilitate se află în zile de naștere pe care modelul, practic, nu le-a văzut — iar o zi de naștere nu are un tipar pe care să te bazezi (nu poți *calcula* ziua

de naștere a cuiva). Așa că o dată văzută o singură dată, sau niciodată, este o monedă pe care modelul nu o poate cântări, și va greși în aproximativ **unul din cinci** cazuri. Nicio ingeniozitate nu repară asta: nu era nimic de învățat.
Acesta este întregul argument în miniatură: rata singletonurilor măsoară cât din lume este neînvățabil *din aceste date*, iar asta devine un **prag** sub erori. Este și motivul pentru care un model aproape nu ratează niciodată un oraș-capitală — *Paris* apare constant, rata lui de singletonuri este aproape zero, deci există destul de mult de învățat.

Asta explică de unde vin halucinațiile. Nu explică de ce *supraviețuiesc* — de ce modelele, după tot antrenamentul ulterior menit să le facă utile și oneste, continuă să blufeze în loc să admită îndoiala. Aici analogia lucrării este aproape incomod de potrivită. Imaginează-ți un student la examen care nu știe răspunsul. Dacă o foaie lăsată goală primește zero, iar o presupunere ar putea primi unu, mișcarea care maximizează nota este să ghicească — cu încredere, specific, niciodată „nu sunt sigur”. Studenții învață asta. Se pare că și modelele o învață — pentru că le notăm la fel. Autorii au trecut prin benchmarkurile pe care domeniul chiar concurează, clasamentele pe care modelele sunt optimizate să le urce, și au găsit că aproape toate dau lui „nu știu” exact același scor ca unui răspuns greșit: zero. Sub această regulă, un model care ghicește mereu va bate un model altfel identic care își marchează onest incertitudinea. Într-un sens destul de literal, le antrenăm prin scor să facă asta.



Benchmarkurile pot face ghicitul rațional. Sub scorarea folosită de majoritatea benchmarkurilor (stânga), un răspuns greșit și un „nu știu” onest primesc ambele zero — deci o presupunere poate doar ajuta. Dacă regulile sunt declarate în întrebare, cu penalizare pentru greșeală (dreapta), abținerea când modelul nu este sigur poate deveni mișcarea mai bună. Asta schimbă ce recompensează testul; nu rezolvă, de una singură, halucinația.

Aceasta este partea care merită păstrată, pentru că merge împotriva titlului obișnuit. Halucinația este adesea vândută ca o limită inevitabilă, aproape mistică, a tehnologiei. Lucrarea contestă ambele idei. Pragul din pretraining nu este un mister — este eroare statistică obișnuită, de tipul pe care machine learning îl înțelege de decenii. Iar persistența nu este inevitabilă — este, în parte, un stimulent pe care l-am construit și pe care l-am putea schimba. Un sistem care ar refuza pur și simplu să răspundă când nu este sigur nu ar halucina deloc; motivul pentru care modelele folosite în practică nu se comportă așa este că scoreboardurile noastre pedepsesc refuzul.

Ce au făcut autorii

Lucrarea are trei părți. Prima, un argument matematic că o anumită cantitate de halucinație este forțată statistic în timpul pretrainingului, arătând că „a genera numai text valid" este cel puțin la fel de dificil ca o problemă binară de clasificare „este această afirmație validă?". A doua, un argument — susținut de o analiză a zece benchmarkuri influente — că metricile principale de tip acuratețe recompensează ghicitul în locul abținerii. A treia, o soluție propusă și un studiu de caz care o testează: evaluări cu **rubrică deschisă**, în care scorarea este declarată chiar în întrebare (de exemplu, „un răspuns corect primește 1, unul greșit −1, deci abține-te dacă ești sigur sub 50%"), astfel încât modelul să poată vedea când onestitatea este recompensată. Ei încearcă asta pe patru modele de frontieră — Gemini 3 Pro de la Google, GPT-5 de la OpenAI, Grok 4 de la xAI și Claude Opus 4.5 de la Anthropic — folosind cele 4.326 de întrebări factuale din SimpleQA. Sunt expliciți că studiul de caz este ilustrativ, „nu o evaluare controlată între modele" (setări implicite, fără tuning, fără normalizare a costurilor).

Ce au găsit

- **Pretrainingul forțează o parte din eroare.** Rata cu care un model emite falsuri încrezătoare este limitată inferior de aproximativ dublul ratei de eroare a celui mai bun clasificator „este această afirmație validă?" construit din el. Pentru fapte fără tipar învățabil, acel prag este cel puțin *rata singletonurilor* — fracțiunea de fapte care apar exact o dată în antrenare. O anumită halucinație este inevitabilă chiar și cu date perfect curate.
- **Scorarea recompensează concret ghicitul.** Sub scorarea obișnuită corect/greșit, a nu te abține niciodată este strategia optimă, iar analiza autorilor găsește că marea majoritate a benchmarkurilor populare notează „nu știu" pur și simplu ca greșit. Un exemplu viu din propriile lor rezultate: pe testul SimpleQA, acuratețea brută îl *favorizează ușor* pe o4-mini de la OpenAI — care răspunde aproape la orice și greșește mai mult de trei sferturi din timp — față de GPT-5-mini, care face mult mai puține greșeli pentru că se abține când nu este sigur. Modelul mai imprudent arată mai bine în clasament.
- **Rubricile deschise inversează stimulentul (în studiul lor de caz).** Ei testează o mitigare simplă a halucinației (modelul răspunde de două ori și se abține dacă cele două răspunsuri nu coincid). Sub acuratețea standard, mitigarea reduce erorile *dar reduce și acuratețea* — deci metrica descurajează adoptarea ei. Sub rubrici deschise, aceeași mitigare iese mai bine pentru toate cele patru modele într-un interval de penalizări; iar GPT-5-mini — pe care acuratețea brută îl penalizase pentru că

se abținea când era nesigur — îl depășește pe o4-mini odată ce scorarea este declarată deschis (n = 4.326 întrebări per model).

Ce înseamnă probabil

Reducerea halucinației nu ține în primul rând de inventarea mai multor teste specifice pentru halucinație. Ține de schimbarea modului în care benchmarkurile principale notează incertitudinea, astfel încât a admite „nu știu” să nu mai fie pedepsit. Până când scoreboardul se schimbă, reducerea halucinației va continua să coste puncte de acuratețe și va continua să fie descurajată — de aceea autorii descriu problema ca „socio-tehnică”: pe de o parte metrică mai bună, pe de altă parte convingerea clasamentelor influente să o adopte.

Ce nu demonstrează

- Nu arată că rubricile deschise repară halucinația în lumea reală. Experimentul de sprijin este un studiu de caz mic și deliberat **necontrolat** — patru modele la setări implicite, o mitigare aleasă, un singur test de întrebări factuale — menit să demonstreze *inversarea stimulentului*, nu să claseze modelele sau să dovedească eficacitate generală.
- Nu susține că scorarea este *singura* cauză. Erorile din datele de antrenare, problemele cu adevărat dificile și prompturile nefamiliare rămân surse separate.
- Nu susține linia populară că halucinațiile sunt inevitabile. Autorii argumentează invers: un sistem care ar răspunde numai la întrebări verificabile și altfel ar spune „nu știu” nu ar halucina niciodată.
- Nu face să dispară pragul din pretraining — îl explică și îl mărginește, iar limita privește erori factuale încrezătoare, nu tot comportamentul modelului.
- Nu arată că rubricile deschise sunt *suficiente* de unele singure. Ele schimbă ce recompensează o evaluare; nu înlocuiesc retrievalul, folosirea de instrumente sau modele mai bine calibrate.

Cât de puternică este evidența?

- Nucleul este **matematic** — limite inferioare formale, nu măsurători. Ca argument teoretic, stă în picioare în propriii termeni.
- Se bazează pe **modele deliberat simplificate** ale problemei; autorii înșiși semnalează „falsa trihotomie” de a trata fiecare răspuns ca fiind corect, incorect sau „nu știu”, și cadrul idealizat al „faptelor arbitrare” folosit pentru limita cea mai curată.
- Analiza benchmarkurilor este un **eșantion mic și selectat** — zece evaluări influente, nu un audit exhaustiv.
- Studiul de caz este **real dar limitat**: patru modele de frontieră, o singură mitigare, numai SimpleQA, setări implicite, explicit „nu o evaluare controlată”. Este o probă de concept pentru argu-

mentul stimulentelor, nu un rezultat de clasament.

- Merită numit punctul de vedere: trei dintre cei patru autori sunt sau au fost angajați ai **OpenAI**, iar lucrarea argumentează că domeniul ar trebui să schimbe felul în care evaluează modelele. Este o poziție bine argumentată din partea unei părți interesate, nu o revizie neutră din exterior — de cântărit, nu de respins. (În favoarea ei, lucrarea își îndreaptă critica spre propriile modele, o4-mini și GPT-5-mini, la fel de direct ca spre altele.)

De ce contează

Reformulează o problemă foarte încărcată de hype. „Halucinația" tinde să fie vândută fie ca un defect straniu, fie ca un zid de netrecut; această lucrare o face obișnuită și parțial auto-produsă — un prag statistic pe care îl putem înțelege, așezat peste un stimulent pe care l-am ales. Lecția mai largă este mai liniștită și mai utilă: progresul în fiabilitate poate depinde la fel de mult de *ce măsurăm* ca de ce construim.

Rezumat curat

Răspunsurile false și încrezătoare ale modelelor lingvistice vin din două surse. Prima este statistică: atunci când un fapt nu are niciun tipar de învățat, un model antrenat să imite limbajul va greși uneori, iar acel prag poate fi estimat (de exemplu, din câte fapte apar o singură dată în antrenare). A doua este stimulentul: aproape fiecare benchmark pe care modelele sunt clasate notează „nu știu" la fel ca un răspuns greșit, așa că ghicitul câștigă mereu — până la punctul în care un model care greșește trei sferturi din timp poate depăși unul mai onest care se abține. Propunerea autorilor nu este încă un test de halucinație, ci „rubrici deschise": declararea scorării în întrebare. Într-un studiu de caz pe patru modele de frontieră, asta inversează stimulentul, astfel încât o metodă care reduce halucinația este recompensată în loc să fie penalizată. Este o lucrare teoretică plus analiză de benchmarkuri plus un experiment mic, explicit necontrolat, peer-reviewed în *Nature*; soluția este promițătoare, dar nu este încă demonstrată larg și la scară, iar halucinațiile sunt prezentate ca nefiind nici misterioase, nici strict inevitabile.

Verificare fără ocolișuri

Ce arată lucrarea: o limită inferioară matematică sub care o parte din halucinație este forțată în timpul pretrainingului (cel puțin „rata singletonurilor" pentru fapte fără tipar); o analiză care găsește că majoritatea benchmarkurilor de vârf nu dau credit pentru „nu știu"; și un studiu de caz pe patru modele în care declararea scorării în prompt („rubrici deschise") face să câștige o metodă de reducere a halucinației, acolo unde acuratețea simplă o penalizase.

Ce este plauzibil dar nu demonstrat: că rubricile deschise, integrate în benchmarkurile principale,

ar reduce semnificativ halucinația în modelele folosite în practică. Experimentul de sprijin este mic și explicit necontrolat.

Ce nu arată: că halucinațiile sunt inevitabile (argumentează invers); că scorarea este singura cauză; că halucinația poate fi eliminată complet; că studiul de caz clasează cele patru modele între ele.

Limitări principale: un model deliberat simplificat corect/incorect/„nu știu” (autorii îl numesc o „falsă trihotomie”); o analiză mică și selectată de benchmarkuri (zece evaluări); un studiu de caz necontrolat (patru modele, o mitigare, un test, setări implicite); și un argument condus de OpenAI despre cum ar trebui domeniul să evalueze modelele.

Câtă încredere ar trebui să aibă un cititor general? Mare că halucinația nu este nici misterioasă, nici strict inevitabilă, și că benchmarkurile mainstream recompensează în prezent ghicitul. Moderată că soluția propusă ajută — are acum o probă de concept reală, dar nu încă o demonstrație că funcționează larg și la scară.

Surse

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