The Clean Paper

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

Hvorfor sprogmodeller hallucinerer — og hvorfor den måde vi bedømmer dem på, fastholder det

De selvsikre usandheder, vi kalder "hallucinationer", er ikke en mystisk fejl: nogle er et statistisk biprodukt af træningen, og de består, fordi gængse benchmarks belønner et selvsikkert gæt frem for et ærligt "jeg ved ikke". Et case study på fire førende modeller viser, at det at angive bedømmelsesreglerne i prompten ("åbne bedømmelsesregler") vender dette incitament om.

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

Hvorfor modeller gætter, og hvorfor vi lærte dem det

Bed en stor sprogmodel om en fremmeds fødselsdag, og den kan svare "7. marts" med samme rolige selvsikkerhed som en, der læser det op fra et kort — og tage fejl, tre gange i træk, med tre forskellige datoer. Forfatterne giver netop dette eksempel: førende modeller, der fik et helt almindeligt faktaspørgsmål — en persons fødselsdag, eller hvad en obskur forkortelse står for — og som hver især selvsikkert fandt på et andet svar, ingen af dem korrekt. Branchens ord for dette er *hallucination*, hvilket får det til at lyde som en fejl i perceptionen. Artiklens første greb er at fjerne mystikken.

Start med, hvordan en model bygges. I sin første og største træningsfase lærer den i praksis, hvordan flydende sprog ser ud, ved at læse en enorm mængde tekst. Tag nu et faktum uden mønster bag sig — én bestemt persons fødselsdag. Hvis den dato optrådte i træningsteksten én gang, eller aldrig, er der intet for en mønstergenkender at gribe fat i: svaret er, set fra modellens synspunkt, vilkårligt. Forfatterne gør dette præcist ved at låne en gammel idé (Alan Turings, fra et helt andet problem): hvis hver femte fødselsdag kun optræder én gang i data, bør en model forventes at tage fejl på mindst hver femte af dem — ikke fordi den er defekt, men fordi der aldrig var noget at lære. (Efter samme logik tager modeller næsten aldrig fejl af et lands hovedstad: de optræder konstant.) De argumenterer, med en vis omhu, for at det i sig selv er et svært problem at skelne et sandt udsagn fra en troværdig løgn, og at det at producere *kun* sande udsagn er mindst lige så svært som det. Et gulv af fejl er indbygget.

Ideen bagved: hvordan man tæller det, man endnu ikke har set

Dette hviler på en genuint smart idé — ældre end sprogmodeller, og værd at stifte ordentligt bekendtskab med.

Start med en pose farvede kugler. Du ved ikke, hvor mange farver den indeholder. Du trækker 100, én ad gangen, og fører regnskab: rød 40, blå 25, grøn 15, gul 5, lilla 3, orange 2 — og derefter **ti forskellige farver, der hver optræder præcis én gang.**

Nu kommer det spørgsmål, Turing faktisk stod over for, i et helt andet problem: *hvor stor er chancen for, at den næste kugle har en farve, du slet ikke har set?* Du kan ikke tælle det, du aldrig har trukket — men du **kan** tælle de farver, du har set præcis én gang, "singletons". Tricket, kaldet **Good–Turing-estimering**, går ud på, at andelen af dine træk, der er singletons, estimerer den sandsynlighed, der stadig gemmer sig i de farver, du ikke har set. Ti af dine hundrede træk var engangsfarver, så chancen for, at den næste kugle har en helt ny farve, er omtrent **10 / 100 = 10 %.**

De farver, du kun har set én gang, er ikke *fejl*. De er et mål for din egen uvidenhed: at mange farver dukker op kun én gang, er stikprøvens måde at fortælle dig på, at verden rummer mere, end du blot endnu ikke har trukket.

Skift nu farver ud med fødselsdage, og posen med modellens træningstekst. Antag, at **hver femte** af de fødselsdage, den har set, forekommer præcis én gang. Samme trick: omtrent en femtedel af sandsynligheden ligger i fødselsdage, modellen reelt aldrig har set — og en fødselsdag har intet mønster at falde tilbage på (man kan ikke *regne sig frem til* nogens fødselsdag). Så en dato set én gang, eller aldrig, er en mønt, modellen ikke kan vægte, og den vil tage fejl på omtrent **hver femte** af dem. Ingen mængde dygtighed retter op på dette: der var aldrig noget at lære.

> Det er hele argumentet i miniature: singleton-frekvensen måler, hvor meget af verden der er ulærbart *ud fra netop disse data*, og det bliver et **gulv** under fejlene. Det er også derfor, en model næsten aldrig tager fejl af en hovedstad — *Paris* optræder konstant, dens singleton-frekvens er tæt på nul, så der er masser at lære.

Det forklarer, hvor hallucinationer kommer fra. Det forklarer ikke, hvorfor de *overlever* — hvorfor modeller, efter al den senere træning, der skal gøre dem hjælpsomme og ærlige, alligevel bluffer frem for at indrømme tvivl. Her er artiklens analogi næsten ubehageligt rammende. Forestil dig en elev til en eksamen, der ikke kender svaret. Hvis et blankt svar giver nul point, og et gæt måske giver ét point, er det pointmaksimerende træk at gætte — selvsikkert, specifikt, aldrig "jeg er ikke sikker". Elever lærer dette. Det viser sig, at modeller gør det samme — fordi vi bedømmer dem på samme måde. Forfatterne gennemgik de benchmarks, feltet faktisk konkurrerer på, de ranglister modeller finjusteres til at klatre på, og fandt, at næsten alle giver "ved ikke" nøjagtig samme score som et forkert svar: nul. Under den regel vinder en model, der altid gætter, over en ellers identisk model, der ærligt flager sin usikkerhed. Vi scorer dem, i temmelig bogstavelig forstand, ind i den adfærd.

Benchmarks kan gøre det rationelt at gætte. Med den scoring, de fleste benchmarks bruger (venstre), giver både et forkert svar og et ærligt "ved ikke" nul point — så et gæt kan kun hjælpe. Angiv reglerne i spørgsmålet, med et fradrag for at tage fejl (højre), og at afstå, når man er usikker, kan blive det bedre træk. Det ændrer, hvad testen belønner; det løser ikke i sig selv hallucination.

Dette er den del, det er værd at huske, for den strider mod den sædvanlige overskrift. Hallucination sælges ofte som en uundgåelig, næsten mystisk begrænsning ved teknologien. Artiklen bestrider

begge dele. Fortræningsgulvet er ikke noget mysterium — det er almindelig statistisk fejl, af den slags maskinlæring har forstået i årtier. Og at det består, er ikke uundgåeligt — det er delvist et incitament, vi selv har bygget, og som vi kan ændre. Et system, der simpelthen undlod at svare, når det var usikkert, ville slet ikke hallucinere; grunden til, at driftsatte modeller ikke opfører sig sådan, er, at vores scoretavler straffer det at afstå.

Hvad forfatterne gjorde

Artiklen har tre dele. Først et matematisk argument for, at noget hallucination er statistisk tvunget frem under fortræningen, ved at vise, at "generér kun gyldig tekst" er mindst lige så svært som et binært klassifikationsproblem af typen "er dette udsagn gyldigt?". For det andet et argument — underbygget af en gennemgang af ti indflydelsesrige benchmarks — for, at gængse nøjagtighedsmål belønner gætteri frem for at afstå. For det tredje et løsningsforslag og en case study, der tester det: evalueringer med **åbne bedømmelsesregler** ("open rubrics"), hvor scoringen angives inde i selve spørgsmålet (for eksempel "et korrekt svar giver 1 point, et forkert −1, så afstå hvis du er mindre end 50 % sikker"), så en model kan afgøre, hvornår ærlighed belønnes. De afprøver dette på fire førende modeller — Googles Gemini 3 Pro, OpenAIs GPT-5, xAIs Grok 4 og Anthropics Claude Opus 4.5 — ved hjælp af SimpleQAs 4.326 faktaspørgsmål. De er tydelige om, at case studyet er illustrativt, "ikke en kontrolleret evaluering på tværs af modeller" (standardindstillinger, ingen finjustering, ingen omkostningsnormalisering).

Hvad de fandt

- **Fortræningen tvinger nogle fejl frem.** Den hastighed, hvormed en model fremsætter selvsikre usandheder, er begrænset nedadtil af (cirka det dobbelte af) fejlraten hos den bedste klassifikator af typen "er dette udsagn gyldigt?", der kan bygges ud fra den. For fakta uden lærbart mønster er det gulv mindst *singleton-frekvensen* — andelen af fakta, der optræder præcis én gang i træningen. Noget hallucination er uundgåeligt selv med perfekt rene data.
- **Bedømmelsen belønner gætteri — konkret.** Under almindelig rigtigt/forkert-scoring er det optimale aldrig at afstå, og forfatternes gennemgang finder, at langt størstedelen af populære benchmarks scorer "ved ikke" som simpelthen forkert. Et sigende eksempel fra deres egen side: på SimpleQA-testen favoriserer den rå nøjagtighed lidt OpenAIs o4-mini — som svarer på næsten alt og tager fejl mere end tre fjerdedele af gangene — frem for GPT-5-mini, som begår langt færre fejl, fordi den afstår, når den er usikker. Den mere hensynsløse model ser bedre ud på scoretavlen.
- **Åbne bedømmelsesregler vender incitamentet om (i deres case study).** De tester et enkelt tiltag mod hallucination (lad modellen svare to gange og afstå, hvis de to svar er uenige). Under standardnøjagtighed reducerer tiltaget fejlene *men reducerer også nøjagtigheden* — så metrikken fraråder at tage det i brug. Under åbne bedømmelsesregler kommer det samme tiltag bedre ud for alle fire modeller på tværs af en række strafniveauer; og GPT-5-mini — som den rå nøjagtighed havde straffet for at afstå, når den var usikker — kommer bedre ud end o4-mini, så snart scoringen angives åbent (n = 4.326 spørgsmål pr. model).

Hvad dette formentlig betyder

At reducere hallucination handler for det meste ikke om at opfinde flere hallucinationsspecifikke tests. Det handler om at ændre, hvordan de gængse benchmarks scorer usikkerhed, så det ikke længere straffer sig at indrømme "ved ikke". Indtil scoretavlen ændres, vil det at reducere hallucination fortsætte med at koste modellerne nøjagtighedspoint og dermed fortsat blive frarådet — hvilket er grunden til, at forfatterne rammer problemet som "socioteknisk": dels en bedre metrik, dels at få de indflydelsesrige ranglister til at tage den i brug.

Hvad dette *ikke* beviser

- Det viser **ikke**, at åbne bedømmelsesregler løser hallucination i den virkelige verden. Det understøttende eksperiment er en lille, bevidst **ukontrolleret** case study — fire modeller med standardindstillinger, ét udvalgt tiltag, én faktaspørgsmålstest — ment til at vise *incitamentsomvendingen*, ikke til at rangere modeller eller bevise generel virkning.
- Den hævder **ikke**, at bedømmelsen er den *eneste* årsag. Fejl i træningsdata, genuint svære problemer og ukendte prompts forbliver separate fejlkilder.
- Den understøtter **ikke** den populære opfattelse, at hallucinationer er uundgåelige. Forfatterne hævder det modsatte: et system, der kun besvarede kontrollerbare spørgsmål og ellers sagde "ved ikke", ville aldrig hallucinere.
- Den får **ikke** fortræningsgulvet til at forsvinde — den forklarer og afgrænser det, og afgrænsningen vedrører selvsikre faktafejl, ikke al modeladfærd.
- Den viser **ikke**, at åbne bedømmelsesregler er *tilstrækkelige* alene. De ændrer, hvad en evaluering belønner; de er ikke en erstatning for informationssøgning, værktøjsbrug eller bedre kalibrerede modeller.

Hvor stærk er evidensen?

- Kernen er **matematik** — formelle nedre grænser, ikke målinger. Som teoretisk argument holder det på egne præmisser.
- Det hviler på **bevidst forenklede modeller** af problemet; forfatterne selv flager den "falske trikotomi" ved at behandle ethvert svar som korrekt, forkert eller "ved ikke", samt det idealiserede scenarie med "vilkårlige fakta", der bruges til den reneste grænse.
- Gennemgangen af benchmarks er et **lille, udvalgt stikprøve** — ti indflydelsesrige evalueringer, ingen udtømmende revision.
- Case studyet er **reelt, men begrænset**: fire førende modeller, ét enkelt tiltag, kun SimpleQA, standardindstillinger, udtrykkeligt "ikke en kontrolleret evaluering". Det er et principbevis for incitamentsargumentet, ikke et benchmarkresultat.
- Værd at nævne udgangspunktet: tre af de fire forfattere er eller har været ansat hos **OpenAI**, og artiklen argumenterer for, at feltet bør ændre, hvordan det evaluerer modeller. Det er en

velargumenteret holdning fra en part med egeninteresse, ikke en neutral ekstern gennemgang — noget at veje ind, ikke afvise. (Til dens ære retter artiklen kritikken mod sine egne modeller, o4-mini og GPT-5-mini, lige så villigt som mod andres.)

Hvorfor det betyder noget

Den omformulerer et kraftigt overhypet problem. "Hallucination" plejer at blive solgt enten som en uhyggelig defekt eller som en umulig mur; denne artikel gør det hverdagsagtigt og delvist selvforskyldt — et statistisk gulv, vi rent faktisk kan forstå, oven på et incitament, vi selv har valgt. Den bredere lære er stille og mere brugbar: yderligere fremskridt inden for pålidelighed kan afhænge lige så meget af *hvad vi måler*, som af hvad vi bygger.

Kort sammenfatning

Selvsikre forkerte svar fra sprogmodeller kommer fra to steder. Det første er statistisk: når et faktum ikke har noget mønster at lære, vil en model, der er trænet til at efterligne sprog, af og til tage fejl, og det gulv kan estimeres (for eksempel ud fra, hvor mange fakta der kun optræder én gang i træningen). Det andet er incitamenter: næsten hvert eneste benchmark, som modeller rangeres på, scorer "ved ikke" på samme måde som et forkert svar, så gætteri vinder altid — i en sådan grad, at en model, der tager fejl tre fjerdedele af gangene, kan slå en mere ærlig model, der afstår. Forfatternes forslag er ikke endnu en hallucinationstest, men "åbne bedømmelsesregler" ("open rubrics"): angiv scoringen inde i spørgsmålet. I en case study med fire førende modeller vender det incitamentet om, så en hallucinationsreducerende metode belønnes i stedet for at blive straffet. Det er en artikel, der kombinerer teori og gennemgang med et lille, udtrykkeligt ukontrolleret eksperiment, fagfællebedømt i *Nature*; løsningen er lovende, men er endnu ikke vist at fungere i stor skala, og hallucinationer argumenteres hverken at være mystiske eller strengt uundgåelige.

No-BS-tjek

Hvad artiklen viser: En matematisk nedre grænse for, at noget hallucination tvinges frem under fortræningen (mindst "singleton-frekvensen" for mønsterløse fakta); en gennemgang, der finder, at de fleste førende benchmarks ikke giver "ved ikke" nogen kredit; samt en case study med fire modeller, hvor det at angive scoringen i spørgsmålet ("åbne bedømmelsesregler") får en hallucinationsreducerende metode til at vinde, hvor ren nøjagtighed havde straffet den.

Hvad der er plausibelt, men ikke bevist: At åbne bedømmelsesregler, indført i de gængse benchmarks, meningsfuldt ville reducere hallucination i driftsatte modeller. Det understøttende eksperiment er lille og udtrykkeligt ukontrolleret.

Hvad den ikke viser: At hallucinationer er uundgåelige (den hævder det modsatte); at bedømmelse er den eneste årsag; at hallucination kan elimineres helt; at case studyet rangerer de fire modeller mod hinanden.

Vigtigste begrænsninger: En bevidst forenklet korrekt/forkert/"ved ikke"-model (forfatterne kalder den en "falsk trikotomi"); en lille, udvalgt gennemgang af benchmarks (ti evalueringer); en ukontrolleret case study (fire modeller, ét tiltag, én test, standardindstillinger); samt et OpenAI-ledet argument om, hvordan feltet bør evaluere modeller.

Hvor stor tillid bør en almindelig læser have? Høj tillid til, at hallucination hverken er mystisk eller strengt uundgåelig, og til, at gængse benchmarks i øjeblikket belønner gætteri. Moderat tillid til, at det foreslåede tiltag hjælper — det har nu et reelt principbevis, men endnu ingen demonstration af, at det virker bredt og i stor skala.

Kilder

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