The Clean Paper

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

Hvorfor språkmodeller hallusinerer — og hvorfor måten vi bedømmer dem på opprettholder det

De selvsikre usannhetene vi kaller «hallusinasjoner», er ikke en mystisk feil: noen er et statistisk biprodukt av treningen, og de består fordi gjengse benchmarker belønner en selvsikker gjetning fremfor et ærlig «jeg vet ikke». En fallstudie på fire ledende modeller viser at det å angi bedømmelsesreglene i spørsmålet («åpne bedømmelsesregler») snur dette insentivet.

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 gjetter, og hvorfor vi lærte dem å gjøre det

Be en stor språkmodell om en fremmeds bursdag, og den kan svare «7. mars» med samme stødige selvsikkerhet som noen som leser det opp fra et kort — og ta feil, tre ganger på rad, med tre ulike datoer. Forfatterne gir nettopp et slikt eksempel: ledende modeller som fikk et helt vanlig faktaspørsmål — en persons bursdag, eller hva en obskur forkortelse står for — og som hver for seg selvsikkert diktet opp et annet svar, ingen av dem riktig. Bransjens ord for dette er *hallusinasjon*, noe som får det til å høres ut som en persepsjonsfeil. Artikkelens første grep er å ta bort mystikken.

Start med hvordan en modell bygges. I sitt første og største treningssteg lærer den, i praksis, hvordan flytende språk ser ut ved å lese en enorm mengde tekst. Ta nå et faktum uten mønster bak seg — én bestemt persons bursdag. Hvis den datoen dukket opp i treningsteksten én gang, eller aldri, er det ingenting for en mønstergjenkjenner å gripe fatt i: svaret er, sett fra modellens ståsted, vilkårlig. Forfatterne gjør dette presist ved å låne en gammel idé (Alan Turings, fra et helt annet problem): hvis hver femte bursdag bare dukker opp én gang i dataene, bør en modell forventes å ta feil på minst hver femte av dem — ikke fordi den er ødelagt, men fordi det aldri fantes noe å lære. (Etter samme logikk tar modeller nesten aldri feil på et lands hovedstad: de dukker opp hele tiden.) De argumenterer, med en viss omhu, for at det å skille et sant utsagn fra en troverdig løgn i seg selv er et vanskelig problem, og at det å produsere *bare* sanne utsagn er minst like vanskelig som det. Et gulv av feil er innebygd.

Idéen bak: hvordan telle det du ennå ikke har sett

Dette hviler på en genuint smart idé — eldre enn språkmodeller, og verdt å bli ordentlig kjent med.

Start med en pose fargede kuler. Du vet ikke hvor mange farger den inneholder. Du trekker 100, én om gangen, og fører regnskap: rød 40, blå 25, grønn 15, gul 5, lilla 3, oransje 2 — og deretter **ti ulike farger som hver dukker opp nøyaktig én gang.**

Nå kommer spørsmålet Turing faktisk sto overfor, i et helt annet problem: *hvor stor er sjansen for at neste kule har en farge du ikke har sett i det hele tatt?* Du kan ikke telle det du aldri har trukket — men du **kan** telle fargene du har sett nøyaktig én gang, «singlene». Trikset, kalt **Good–Turing-estimering**, går ut på at andelen av trekningene dine som er singler, estimerer sannsynligheten som fremdeles skjuler seg i fargene du ikke har sett. Ti av de hundre trekningene dine var engangsfarger, så sjansen for at neste kule har en helt ny farge, er omtrent **10 / 100 = 10 %.**

De en gang sette fargene er ikke *feil*. De er et mål på din egen uvitenhet: at mange farger dukker opp bare én gang, er stikkprøvens måte å fortelle deg at verden rommer mer enn du rett og slett ikke har trukket ennå.

Bytt nå ut farger med bursdager, og posen med modellens treningstekst. Anta at **hver femte** av bursdagene den har sett, forekommer nøyaktig én gang. Samme triks: omtrent en femtedel av sannsynligheten ligger i bursdager modellen i praksis aldri har sett — og en bursdag har ikke noe mønster å falle tilbake på (man kan ikke *regne seg frem* til noens bursdag). Så en dato sett én gang, eller aldri, er et mynt modellen ikke kan vekte, og den vil ta feil på omtrent **hver**

> **femte** av dem. Ingen mengde dyktighet retter opp dette: det fantes aldri noe å lære.
> Det er hele argumentet i miniatyr: singel-frekvensen måler hvor mye av verden som er ulærbart *fra nettopp disse dataene*, og det blir et **gulv** under feilene. Det er også derfor en modell nesten aldri bommer på en hovedstad — *Paris* dukker opp hele tiden, singel-frekvensen er nær null, så det er nok å lære.

Det forklarer hvor hallusinasjoner kommer fra. Det forklarer ikke hvorfor de *overlever* — hvorfor modeller, etter all den senere treningen som skal gjøre dem hjelpsomme og ærlige, likevel bløffer heller enn å innrømme tvil. Her er artikkelens analogi nesten ubehagelig treffende. Se for deg en elev til eksamen som ikke kan svaret. Hvis et blankt svar gir null poeng og en gjetning kan gi ett poeng, er det poengmaksimerende trekket å gjette — selvsikkert, spesifikt, aldri «jeg er ikke sikker». Elever lærer dette. Det viser seg at modeller gjør det også — fordi vi bedømmer dem på samme måte. Forfatterne gikk gjennom de benchmarkene feltet faktisk konkurrerer på, de rangeringslistene modeller finjusteres for å klatre på, og fant at nesten alle gir «vet ikke» nøyaktig samme poengsum som et feil svar: null. Under den regelen vinner en modell som alltid gjetter, over en for øvrig identisk modell som ærlig flagger sin usikkerhet. Vi poengsetter dem, i temmelig bokstavelig forstand, inn i den atferden.

Benchmarker kan gjøre det rasjonelt å gjette. Med poengsettingen de fleste benchmarker bruker (venstre), gir både et feil svar og et ærlig «vet ikke» null poeng — så en gjetning kan bare hjelpe. Angi reglene i spørsmålet, med et trekk for å ta feil (høyre), og det å avstå når man er usikker kan bli det bedre trekket. Det endrer hva testen belønner; det løser ikke hallusinasjon i seg selv.

Dette er delen som er verdt å huske, for den går imot det vanlige oppslaget. Hallusinasjon selges ofte

som en uunngåelig, nesten mystisk begrensning ved teknologien. Artikkelen bestrider begge deler. Fortreningsgulvet er ikke noe mysterium — det er ordinær statistisk feil, av det slaget maskinlæring har forstått i tiår. Og at det består, er ikke uunngåelig — det er delvis et insentiv vi selv har bygget og kan endre. Et system som rett og slett avsto fra å svare når det var usikkert, ville ikke hallusinere i det hele tatt; grunnen til at driftsatte modeller ikke oppfører seg slik, er at poengtavlene våre straffer det å avstå.

Hva forfatterne gjorde

Artikkelen har tre deler. Først et matematisk argument for at noe hallusinasjon er statistisk tvunget frem under fortreningen, ved å vise at «generer bare gyldig tekst» er minst like vanskelig som et binært klassifiseringsproblem av typen «er dette utsagnet gyldig?». For det andre et argument — underbygget av en gjennomgang av ti innflytelsesrike benchmarker — for at gjengse treffsikkerhetsmål belønner gjetting fremfor det å avstå. For det tredje et løsningsforslag og en fallstudie som tester det: vurderinger med **åpne bedømmelsesregler** («open rubrics»), der poengsettingen angis inne i selve spørsmålet (for eksempel «et korrekt svar gir 1 poeng, et feil svar −1, så avstå hvis du er mindre enn 50 % sikker»), slik at en modell kan avgjøre når ærlighet belønnes. De prøver dette på fire ledende modeller — Googles Gemini 3 Pro, OpenAIs GPT-5, xAIs Grok 4 og Anthropics Claude Opus 4.5 — ved hjelp av SimpleQAs 4 326 faktaspørsmål. De er tydelige på at fallstudien er illustrerende, «ikke en kontrollert evaluering på tvers av modeller» (standardinnstillinger, ingen finjustering, ingen kostnadsnormalisering).

Hva de fant

- **Fortreningen tvinger frem noen feil.** Hastigheten en modell ytrer selvsikre usannheter med, er begrenset nedad av (omtrent det dobbelte av) feilraten til den beste klassifisereren av typen «er dette utsagnet gyldig?» som kan bygges ut fra den. For fakta uten lærbart mønster er det gulvet minst *singel-frekvensen* — andelen fakta som forekommer nøyaktig én gang i treningen. Noe hallusinasjon er uunngåelig selv med perfekt rene data.
- **Bedømmelsen belønner gjetting — konkret.** Ved vanlig riktig/feil-poengsetting er det optimale å aldri avstå, og forfatternes gjennomgang finner at det store flertallet av populære benchmarker poengsetter «vet ikke» som rett og slett feil. Et talende eksempel fra deres eget hold: på SimpleQA-testen favoriserer den rå treffsikkerheten lett OpenAIs o4-mini — som svarer på nesten alt og tar feil mer enn tre firedeler av gangene — fremfor GPT-5-mini, som gjør langt færre feil fordi den avstår når den er usikker. Den mer hensynsløse modellen ser bedre ut på poengtavlen.
- **Åpne bedømmelsesregler snur insentivet (i deres fallstudie).** De tester et enkelt tiltak mot hallusinasjon (la modellen svare to ganger og avstå hvis de to svarene er uenige). Under vanlig treffsikkerhet reduserer tiltaket feilene *men reduserer også treffsikkerheten* — så metrikken frarår å ta det i bruk. Under åpne bedømmelsesregler kommer det samme tiltaket bedre ut for alle fire modellene, over en rekke ulike straffenivåer; og GPT-5-mini — som den rå treffsikkerheten hadde straffet for å avstå når den var usikker — kommer bedre ut enn o4-mini så snart poengsettingen

oppgis åpent (n = 4 326 spørsmål per modell).

Hva dette trolig betyr

Å redusere hallusinasjon handler for det meste ikke om å finne opp flere hallusinasjonsspesifikke tester. Det handler om å endre hvordan de gjengse benchmarkene poengsetter usikkerhet, slik at det å innrømme «vet ikke» ikke lenger straffes. Inntil poengtavlen endres, vil det å redusere hallusinasjon fortsette å koste modellene treffsikkerhetspoeng og dermed fortsette å bli frarådet — noe som er hvorfor forfatterne rammer inn problemet som «sosioteknisk»: dels et bedre mål, dels det å få de innflytelsesrike rangeringslistene til å ta det i bruk.

Hva dette *ikke* beviser

- Det viser **ikke** at åpne bedømmelsesregler løser hallusinasjon i den virkelige verden. Det støttende eksperimentet er en liten, bevisst **ukontrollert** fallstudie — fire modeller med standardinnstillinger, ett valgt tiltak, én faktaspørsmålstest — ment å vise *insentivsnuen*, ikke å rangere modeller eller bevise generell virkningsgrad.
- Den hevder **ikke** at bedømmelsen er den *eneste* årsaken. Feil i treningsdataene, genuint vanskelige problemer og ukjente spørsmål forblir egne feilkilder.
- Den støtter **ikke** den populære oppfatningen om at hallusinasjoner er uunngåelige. Forfatterne hevder det motsatte: et system som bare besvarte kontrollerbare spørsmål og ellers sa «vet ikke», ville aldri hallusinere.
- Den får **ikke** fortreningsgulvet til å forsvinne — den forklarer og avgrenser det, og avgrensningen gjelder selvsikre faktafeil, ikke all modellatferd.
- Den viser **ikke** at åpne bedømmelsesregler er *tilstrekkelige* alene. De endrer hva en evaluering belønner; de er ingen erstatning for informasjonsinnhenting, verktøybruk eller bedre kalibrerte modeller.

Hvor sterke er bevisene?

- Kjernen er **matematikk** — formelle nedre grenser, ikke målinger. Som teoretisk argument holder det på egne premisser.
- Det hviler på **bevisst forenklede modeller** av problemet; forfatterne selv flagger den «falske trikotomien» ved å behandle hvert svar som korrekt, feil eller «vet ikke», samt det idealiserte scenariet med «vilkårlige fakta» som brukes for den reneste grensen.
- Gjennomgangen av benchmarker er et **lite, håndplukket utvalg** — ti innflytelsesrike evalueringer, ingen uttømmende revisjon.
- Fallstudien er **reell, men begrenset**: fire ledende modeller, ett enkelt tiltak, kun SimpleQA, stan-

dardinnstillinger, uttrykkelig «ikke en kontrollert evaluering». Det er et prinsippbevis for insentivargumentet, ikke et benchmarkresultat.

- Verdt å nevne ståstedet: tre av de fire forfatterne er eller har vært ansatt hos **OpenAI**, og artikkelen argumenterer for at feltet bør endre hvordan det evaluerer modeller. Det er et velargumentert standpunkt fra en part med egeninteresse, ingen nøytral ekstern gjennomgang — noe å veie inn, ikke avfeie. (Til dens fortjeneste retter artikkelen kritikken mot sine egne modeller, o4-mini og GPT-5-mini, like villig som mot andres.)

Hvorfor det betyr noe

Den omformulerer et sterkt oppblåst problem. «Hallusinasjon» pleier å bli solgt enten som en skummel defekt eller som en ubevegelig mur; denne artikkelen gjør det hverdagslig og delvis selvpåført — et statistisk gulv vi faktisk kan forstå, oppå et insentiv vi selv har valgt. Den videre lærdommen er stillere og mer nyttig: videre fremskritt på pålitelighet kan avhenge like mye av *hva vi måler* som av hva vi bygger.

Kort sammendrag

Selvsikre feilaktige svar fra språkmodeller kommer fra to steder. Det første er statistisk: når et faktum mangler mønster å lære, vil en modell trent til å etterligne språk av og til ta feil, og det gulvet kan estimeres (for eksempel ut fra hvor mange fakta som bare forekommer én gang i treningen). Det andre er insentiver: nesten hver benchmark modeller rangeres på, poengsetter «vet ikke» likt med et feil svar, så gjetting vinner alltid — i den grad at en modell som tar feil tre firedeler av gangene, kan slå en mer ærlig modell som avstår. Forfatternes forslag er ikke enda en hallusinasjonstest, men «åpne bedømmelsesregler» («open rubrics»): angi poengsettingen inne i spørsmålet. I en fallstudie på fire ledende modeller snur det insentivet slik at en hallusinasjonsreduserende metode belønnes i stedet for å straffes. Det er en artikkel som kombinerer teori og gjennomgang med et lite, uttrykkelig ukontrollert eksperiment, fagfellevurdert i *Nature*; løsningen er lovende, men er ennå ikke vist å fungere i stor skala, og hallusinasjoner argumenteres for verken å være mystiske eller strengt uunngåelige.

No-BS-sjekk

Hva artikkelen viser: En matematisk nedre grense for at noe hallusinasjon tvinges frem under fortreningen (minst «singel-frekvensen» for mønsterløse fakta); en gjennomgang som finner at de fleste ledende benchmarkene ikke gir «vet ikke» noen kreditt; og en fallstudie med fire modeller der det å angi poengsettingen i spørsmålet («åpne bedømmelsesregler») får en hallusinasjonsreduserende metode til å vinne, der ren treffsikkerhet hadde straffet den.

Hva som er plausibelt, men ikke bevist: At åpne bedømmelsesregler, innført i gjengse benchmarker, ville redusere hallusinasjon merkbart i driftsatte modeller. Det støttende eksperimentet er lite og uttrykkelig ukontrollert.

Hva den ikke viser: At hallusinasjoner er uunngåelige (den hevder det motsatte); at bedømmelse er den eneste årsaken; at hallusinasjon kan elimineres helt; at fallstudien rangerer de fire modellene mot hverandre.

Hovedbegrensninger: En bevisst forenklet korrekt/feil/«vet ikke»-modell (forfatterne kaller den en «falsk trikotomi»); en liten, håndplukket gjennomgang av benchmarker (ti evalueringer); en ukontrollert fallstudie (fire modeller, ett tiltak, én test, standardinnstillinger); samt et OpenAI-ledet argument om hvordan feltet bør evaluere modeller.

Hvor mye tillit bør en allmenn leser ha? Høy tillit til at hallusinasjon verken er mystisk eller strengt uunngåelig, og til at gjengse benchmarker for tiden belønner gjetting. Moderat tillit til at det foreslåtte tiltaket hjelper — det har nå et reelt prinsippbevis, men ennå ingen demonstrasjon av at det fungerer 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