The Clean Paper

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

Varför språkmodeller hallucinerar — och varför sättet vi bedömer dem på vidmakthåller det

De självsäkra osanningarna vi kallar ”hallucinationer” är inget mystiskt fel: en del är en statistisk biprodukt av träningen, och de består för att gängse benchmarks belönar en självsäker gissning framför ett ärligt ”jag vet inte”. En fallstudie på fyra ledande modeller visar att det att ange bedömningsreglerna i frågan (”öppna bedömningsregler”) vänder om det incitamentet.

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

Varför modeller gissar, och varför vi lärde dem att göra det

Be en stor språkmodell om en främlings födelsedag, och den kan svara ”7 mars” med samma stadiga säkerhet som någon som läser innantill på ett kort — och ha fel, tre gånger i rad, med tre olika datum. Författarna ger just detta slags exempel: ledande modeller som fick en helt vanlig sakfråga — en persons födelsedag, eller vad en obskyr förkortning står för — och som var för sig självsäkert hittade på ett annat svar, inget av dem korrekt. Branschens ord för detta är *hallucination*, vilket får det att låta som ett fel i perceptionen. Artikelns första drag är att ta bort mystiken.

Börja med hur en modell byggs. I sitt första och största träningssteg lär den sig, i praktiken, hur flytande språk ser ut genom att läsa en enorm mängd text. Ta nu ett faktum utan mönster bakom sig — en viss persons födelsedag. Om det datumet förekom i träningstexten en gång, eller aldrig, finns det inget för en mönsterinlärare att ta fasta på: svaret är, från modellens synvinkel, godtyckligt. Författarna gör detta precist genom att låna en gammal idé (Alan Turings, från ett annat problem): om var femte födelsedag bara förekommer en gång i data, bör en modell förväntas ha fel på minst var femte av dem — inte för att den är trasig, utan för att det aldrig fanns något att lära sig. (Av samma logik har modeller nästan aldrig fel på ett lands huvudstad: de förekommer ständigt.) De argumenterar, med viss omsorg, för att det i sig är ett svårt problem att skilja ett sant påstående från en trovärdig lögn, och att det att producera *enbart* sanna påståenden är minst lika svårt som det. Ett golv av fel är inbyggt.

Idén bakom: hur man räknar det man ännu inte sett

Detta vilar på en genuint smart idé — äldre än språkmodeller, och värd att bekanta sig med ordentligt.

Börja med en påse färgade kulor. Du vet inte hur många färger den innehåller. Du drar 100 kulor, en i taget, och för bok: röd 40, blå 25, grön 15, gul 5, lila 3, orange 2 — och sedan **tio olika färger som var och en dyker upp exakt en gång.**

Nu kommer den fråga Turing faktiskt stod inför, i ett helt annat problem: *hur stor är chansen att nästa kula har en färg du inte sett alls?* Du kan inte räkna det du aldrig dragit — men du **kan** räkna de färger du sett exakt en gång, ”singletons”. Tricket, som kallas **Good–Turing-skattning**, går ut på att andelen av dina dragningar som är singletons skattar den sannolikhet som fortfarande döljer sig i de färger du inte sett. Tio av dina hundra dragningar var engångsfärger, så chansen att nästa kula har en helt ny färg är ungefär **10 / 100 = 10 %.**

De en gång sedda färgerna är inte *misstag*. De är ett mått på din egen okunskap: att många färger dyker upp bara en gång är stickprovets sätt att säga att världen rymmer mer som du helt enkelt ännu inte har dragit.

Byt nu ut färger mot födelsedagar, och påsen mot modellens träningstext. Anta att **var femte** av de födelsedagar den sett förekommer exakt en gång. Samma trick: ungefär en femtedel av sannolikheten finns i födelsedagar modellen i praktiken aldrig har sett — och en födelsedag har inget mönster att falla tillbaka på (man kan inte *räkna ut* någons födelsedag). Ett datum som setts en gång, eller aldrig, är alltså ett mynt modellen inte kan vikta, och den kommer att ha fel på ungefär **var femte** av dem. Ingen mängd skicklighet rättar till detta: det fanns aldrig något att lära sig.

> Det är hela argumentet i miniatyr: singleton-frekvensen mäter hur mycket av världen som är olärbar *från just den här datan*, och det blir ett **golv** under felen. Det är också därför en modell nästan aldrig missar en huvudstad — *Paris* förekommer ständigt, dess singleton-frekvens är nära noll, så det finns gott om att lära.

Det förklarar var hallucinationer kommer ifrån. Det förklarar inte varför de *överlever* — varför modeller, efter all senare träning som är tänkt att göra dem hjälpsamma och ärliga, ändå bluffar hellre än att medge tvivel. Här är artikelns analogi nästan obekvämt träffande. Föreställ dig en elev i en tentamen som inte kan svaret. Om ett tomt svar ger noll poäng och en gissning kan ge en poäng, är det poängmaximerande draget att gissa — självsäkert, specifikt, aldrig ”jag är inte säker”. Elever lär sig detta. Det visar sig att modeller gör det också — eftersom vi bedömer dem på samma sätt. Författarna gick igenom de benchmarks fältet faktiskt tävlar på, de topplistor modeller finjusteras för att klättra på, och fann att nästan alla ger ”vet inte” exakt samma poäng som ett felaktigt svar: noll. Under den regeln vinner en modell som alltid gissar över en i övrigt identisk modell som ärligt flaggar sin osäkerhet. Vi poängsätter dem, i ganska bokstavlig mening, in i det beteendet.

Benchmarks kan göra det rationellt att gissa. Med den poängsättning de flesta benchmarks använder (vänster) ger både ett fel svar och ett ärligt ”vet inte” noll poäng — så en gissning kan bara hjälpa. Ange reglerna i frågan, med ett avdrag för att ha fel (höger), och att avstå när man är osäker kan bli det bättre draget. Det ändrar vad testet belönar; det löser inte i sig hallucination.

Det här är den del som är värd att minnas, för den går emot den vanliga rubriken. Hallucination säljs ofta som en oundviklig, nästan mystisk begränsning i tekniken. Artikeln bestrider båda delarna.

Förträningsgolvet är inget mysterium — det är vanligt statistiskt fel, av det slag maskininlärning har förstått i decennier. Och att det består är inte oundvikligt — det är delvis ett incitament vi själva byggt och kan ändra. Ett system som helt enkelt avstod från att svara när det var osäkert skulle inte hallucinera alls; anledningen till att driftsatta modeller inte beter sig så är att våra poängtavlor bestraffar att avstå.

Vad författarna gjorde

Artikeln har tre delar. Först ett matematiskt argument för att viss hallucination är statistiskt tvingad under förträningen, genom att visa att ”generera enbart giltig text” är minst lika svårt som ett binärt klassificeringsproblem av typen ”är detta påstående giltigt?”. För det andra ett argument — underbyggt av en genomgång av tio inflytelserika benchmarks — för att gängse träffsäkerhetsmått belönar gissande framför att avstå. För det tredje ett förslag till lösning och en fallstudie som testar den: utvärderingar med **öppna bedömningsregler** (”open rubrics”), där poängsättningen anges direkt i frågan (till exempel ”ett korrekt svar ger 1 poäng, ett fel svar −1, så avstå om du är mindre än 50 % säker”), så att en modell kan avgöra när ärlighet belönas. De testar detta på fyra ledande modeller — Googles Gemini 3 Pro, OpenAIs GPT-5, xAI:s Grok 4 och Anthropics Claude Opus 4.5 — med hjälp av SimpleQAs 4 326 sakfrågor. De är tydliga med att fallstudien är illustrativ, ”inte en kontrollerad utvärdering mellan modeller” (standardinställningar, ingen finjustering, ingen kostnadsnormalisering).

Vad de fann

- **Förträningen tvingar fram vissa fel.** Den hastighet med vilken en modell yttrar självsäkra osanningar är begränsad nedåt av (ungefär det dubbla av) felfrekvensen hos den bästa klassificerare av typen ”är detta påstående giltigt?” som kan byggas utifrån den. För fakta utan lärbart mönster är det golvet minst *singleton-frekvensen* — andelen fakta som förekommer exakt en gång i träningen. Viss hallucination är oundviklig även med perfekt rena data.
- **Bedömningen belönar gissande — konkret.** Vid vanlig rätt/fel-poängsättning är det optimala att aldrig avstå, och författarnas genomgång visar att de allra flesta populära benchmarks poängsätter ”vet inte” som helt enkelt fel. Ett talande exempel från deras eget håll: på SimpleQA-testet gynnar den råa träffsäkerheten lätt OpenAIs o4-mini — som svarar på nästan allt och har fel mer än tre fjärdedelar av gångerna — framför GPT-5-mini, som gör betydligt färre misstag eftersom den avstår när den är osäker. Den mer hänsynslösa modellen ser bättre ut på poängtavlan.
- **Öppna bedömningsregler vänder om incitamentet (i deras fallstudie).** De testar en enkel åtgärd mot hallucination (låta modellen svara två gånger och avstå om de två svaren skiljer sig åt). Under vanlig träffsäkerhet minskar åtgärden felen *men minskar också träffsäkerheten* — så måttet avråder från att använda den. Under öppna bedömningsregler ligger samma åtgärd före för alla fyra modellerna över en rad olika straffnivåer, och GPT-5-mini — som den råa träffsäkerheten hade straffat för att avstå när den var osäker — hamnar före o4-mini så snart poängsättningen anges öppet (n = 4 326 frågor per modell).

Vad detta troligen betyder

Att minska hallucination handlar mestadels inte om att uppfinna fler hallucinationsspecifika tester. Det handlar om att ändra hur de gängse benchmarks poängsätter osäkerhet, så att det inte längre straffar sig att medge ”vet inte”. Så länge poängtavlan inte ändras kommer det att minska hallucination att fortsätta kosta modellerna träffsäkerhetspoäng och därmed fortsätta avrådas ifrån — vilket är varför författarna beskriver problemet som ”socioteknisk”: dels ett bättre mått, dels att få de inflytelserika topplistorna att ta det till sig.

Vad detta *inte* bevisar

- Det visar **inte** att öppna bedömningsregler löser hallucination i verkligheten. Det stödjande experimentet är en liten, medvetet **okontrollerad** fallstudie — fyra modeller med standardinställningar, en vald åtgärd, ett sakfrågetest — avsedd att visa *incitamentsomvändningen*, inte att rangordna modeller eller bevisa allmän verkningsgrad.
- Den hävdar **inte** att bedömningen är den *enda* orsaken. Fel i träningsdatan, genuint svåra problem och obekanta frågor förblir egna felkällor.
- Den stöder **inte** den populära uppfattningen att hallucinationer är oundvikliga. Författarna hävdar motsatsen: ett system som bara besvarade kontrollerbara frågor och i övrigt sa ”vet inte” skulle aldrig hallucinera.
- Den får **inte** förträningsgolvet att försvinna — den förklarar och avgränsar det, och avgränsningen gäller självsäkra sakfel, inte allt modellbeteende.
- Den visar **inte** att öppna bedömningsregler är *tillräckliga* på egen hand. De ändrar vad en utvärdering belönar; de är inget substitut för informationssökning, verktygsanvändning eller bättre kalibrerade modeller.

Hur starka är beläggen?

- Kärnan är **matematik** — formella nedre gränser, inte mätningar. Som teoretiskt argument håller det på egna villkor.
- Det vilar på **medvetet förenklade modeller** av problemet; författarna själva flaggar för den ”falska trikotomin” att behandla varje svar som korrekt, felaktigt eller ”vet inte”, samt det idealiserade scenariot med ”godtyckliga fakta” som används för den renaste gränsen.
- Genomgången av benchmarks är ett **litet, handplockat urval** — tio inflytelserika utvärderingar, ingen uttömmande granskning.
- Fallstudien är **verklig men begränsad**: fyra ledande modeller, en enda åtgärd, enbart SimpleQA, standardinställningar, uttryckligen ”inte en kontrollerad utvärdering”. Det är ett principbevis för incitamentsargumentet, inte ett benchmarkresultat.
- Värt att nämna utgångspunkten: tre av de fyra författarna är eller har varit anställda på **OpenAI**, och artikeln argumenterar för att fältet bör ändra hur det utvärderar modeller. Det är en välargu-

menterad ståndpunkt från en part med egenintresse, ingen neutral extern granskning — något att väga in, inte avfärda. (Till dess heder riktar artikeln kritiken mot sina egna modeller, o4-mini och GPT-5-mini, lika villigt som mot andras.)

Varför det spelar roll

Den omformulerar ett kraftigt uppblåst problem. ”Hallucination” brukar säljas antingen som en kuslig defekt eller som en omöjlig mur; den här artikeln gör det vardagligt och delvis självförvållat — ett statistiskt golv vi faktiskt kan förstå, ovanpå ett incitament vi själva valt. Den vidare lärdomen är tystare och mer användbar: fortsatta framsteg för tillförlitlighet kan bero lika mycket på *vad vi mäter* som på vad vi bygger.

Ren sammanfattning

Självsäkra felaktiga svar från språkmodeller kommer från två håll. Det första är statistiskt: när ett faktum saknar mönster att lära sig kommer en modell som är tränad att härma språk ibland att ha fel, och det golvet går att uppskatta (till exempel utifrån hur många fakta som bara förekommer en gång i träningen). Det andra är incitament: nästan varje benchmark som modeller rangordnas på poängsätter ”vet inte” likadant som ett fel svar, så gissning vinner alltid — till den grad att en modell som har fel tre fjärdedelar av gångerna kan slå en mer ärlig modell som avstår. Författarnas förslag är inte ännu ett hallucinationstest utan öppna bedömningsregler (”open rubrics”): ange poängsättningen inne i frågan. I en fallstudie på fyra ledande modeller vänder det om incitamentet så att en hallucinationsminskande metod belönas i stället för att straffas. Det är en artikel som kombinerar teori och genomgång med ett litet, uttryckligen okontrollerat experiment, granskad av fackkolleger i *Nature*; lösningen är lovande men har ännu inte visats fungera i stor skala, och hallucinationer argumenteras varken vara mystiska eller strikt oundvikliga.

No-BS-kontroll

Vad artikeln visar: En matematisk nedre gräns för att viss hallucination tvingas fram under förträningen (minst ”singleton-frekvensen” för mönsterlösa fakta); en genomgång som visar att de flesta ledande benchmarks inte ger ”vet inte” någon poäng alls; samt en fallstudie med fyra modeller där det att ange poängsättningen i frågan (”öppna bedömningsregler”) gör att en hallucinationsminskande metod vinner, där ren träffsäkerhet hade straffat den.

Vad som är rimligt men inte bevisat: Att öppna bedömningsregler, införda i de gängse benchmarksen, skulle minska hallucination i driftsatta modeller på ett betydande sätt. Det stödjande experimentet är litet och uttryckligen okontrollerat.

Vad den inte visar: Att hallucinationer är oundvikliga (den hävdar motsatsen); att bedömningen är den enda orsaken; att hallucination kan elimineras helt; att fallstudien rangordnar de fyra modellerna mot varandra.

Huvudsakliga begränsningar: En medvetet förenklad modell med korrekt/felaktigt/”vet inte” (författarna kallar den en ”falsk trikotomi”); en liten, handplockad genomgång av benchmarks (tio utvärderingar); en okontrollerad fallstudie (fyra modeller, en åtgärd, ett test, standardinställningar); samt ett OpenAI-lett argument om hur fältet bör utvärdera modeller.

Hur stor tilltro bör en allmän läsare ha? Stor tilltro till att hallucination varken är mystiskt eller strikt oundvikligt, och till att gängse benchmarks för närvarande belönar gissande. Måttlig tilltro till att den föreslagna lösningen hjälper — den har nu ett verkligt principbevis, men ännu ingen demonstration av att den fungerar brett och i stor skala.

Källor

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