The Clean Paper

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

Pse modelet gjuhësore halucinojnë — dhe pse mënyra si i vlerësojmë i shtyn të vazhdojnë

Pohimet e rreme me vetëbesim që quajmë "halucinacione" nuk janë defekt misterioz: disa janë nënprodukt statistikor i trajnimit dhe vazhdojnë sepse benchmark-et kryesore shpërblejnë një hamendësim me vetëbesim më shumë se një "nuk e di" të ndershëm. Një case study me katër modele frontier tregon se deklarimi i rregullave të pikëzimit në prompt ("open rubrics") e përmbys këtë stimul.

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

Pse modelet hamendësojnë — dhe pse i kemi mësuar ta bëjnë

Pyet një model të madh gjuhësor për ditëlindjen e një personi të panjohur dhe mund të të thotë "7 mars" me qetësinë e dikujt që po e lexon nga një kartë — dhe të jetë gabim, tri herë radhazi, me tri data të ndryshme. Autorët japin pikërisht shembuj të tillë: modele kryesore që pyeten për një fakt të thjeshtë — ditëlindjen e një personi ose kuptimin e një shkurtese të rrallë — dhe secili shpik me vetëbesim një përgjigje tjetër, asnjëra e saktë. Industria e quan këtë *hallucination*, sikur të ishte defekt perceptimi. Hapi i parë i punimit është t'i heqë misterin.

Fillojmë nga mënyra si ndërtohet një model. Në fazën e parë dhe më të madhe të trajnimit, ai mëson në thelb se si duket gjuha e rrjedhshme duke lexuar sasi të mëdha teksti. Tani merr një fakt pa model të brendshëm — ditëlindjen e një personi të caktuar. Nëse ajo datë është shfaqur vetëm një herë në të dhënat e trajnimit, ose aspak, një sistem që mëson modele nuk ka shumë çfarë të kapë: nga këndvështrimi i tij, përgjigjja është arbitrare. Autorët e bëjnë këtë ide të saktë duke huazuar një teknikë të vjetër, të lidhur me Alan Turing-un në një problem krejt tjetër: nëse një në pesë ditëlindje shfaqet vetëm një herë në të dhëna, duhet pritur që modeli të gabojë të paktën rreth një në pesë të tilla — jo sepse është "i prishur", por sepse nuk kishte mjaftueshëm për të mësuar. (Për të njëjtën arsye modelet pothuajse nuk gabojnë kryeqytetet e shteteve: ato përsëriten vazhdimisht.) Autorët argumentojnë me kujdes se të dallosh një deklaratë të vërtetë nga një të rreme por të besueshme është vetë problem i vështirë dhe se të prodhosh **vetëm** deklarata të vërteta është të paktën po aq e vështirë. Ekziston një dysheme statistikore gabimesh.

Ideja poshtë kësaj: si numërohet ajo që ende nuk e ke parë

Kjo mbështetet në një ide vërtet të zgjuar — më e vjetër se modelet gjuhësore dhe që ia vlen të njihet siç duhet.

Mendo një qese me topa me ngjyra. Nuk e di sa ngjyra ka brenda. Nxjerr 100 topa, një nga një, dhe i numëron: të kuq 40, blu 25, të gjelbër 15, të verdhë 5, vjollcë 3, portokalli 2 — dhe pastaj **dhjetë ngjyra të ndryshme që shfaqen vetëm nga një herë**.

Pyetja që Turing-u përballoi në një problem krejt tjetër ishte: *sa është probabiliteti që topi tjetër të ketë një ngjyrë që nuk e ke parë fare?* Nuk mund të numërosh drejtpërdrejt atë që nuk ke nxjerrë kurrë — por **mund** të numërosh ngjyrat që i ke parë vetëm një herë, "singleton"-ët. Truku, i quajtur **vlerësim Good–Turing**, thotë se përqindja e tërheqjeve që janë singleton vlerëson masën e probabilitetit që fshihet te ngjyrat ende të paparë. Dhjetë nga njëqind tërheqjet e tua ishin ngjyra që u shfaqën vetëm një herë, prandaj probabiliteti që topi tjetër të ketë ngjyrë krejt të re është rreth **10 / 100 = 10%**.

Ngjyrat e shfaqura vetëm një herë nuk janë **gabime**. Janë matje e padijes sonë: shumë ngjyra që dalin vetëm një herë janë mënyra e mostrës për të thënë se bota përmban edhe të tjera që thjesht nuk i kemi parë ende.

Tani zëvendëso ngjyrat me ditëlindje dhe qesen me tekstin e trajnimit. Supozo se, mes ditëlindjeve që modeli ka parë, **një në pesë shfaqet saktësisht një herë**. I njëjti truk: afërsisht një e pesta e probabilitetit qëndron te ditëlindje që modeli në praktikë nuk i ka parë mjaftueshëm — dhe ditëlindja nuk ka model nga i cili mund të nxirret (nuk mund ta **llogaritësh**

> ditëlindjen e dikujt). Pra një datë e parë një herë ose kurrë është një fakt që modeli nuk mund ta inferojë mirë dhe do të gabojë në një pjesë të konsiderueshme të tyre. Asnjë "zgjuarsi" nuk e kompenson plotësisht faktin që informacioni nuk ishte aty.
> Ky është argumenti në miniaturë: norma e singleton-ëve mat sa nga bota mbetet e pamë-sueshme **nga këto të dhëna**, dhe kjo vendos një **kufi të poshtëm** për gabimin. Është edhe arsyeja pse modeli pothuajse nuk gabon kryeqytete — *Paris* shfaqet vazhdimisht, norma singleton është afër zeros dhe ka shumë për të mësuar.

Kjo shpjegon nga vijnë halucinacionet. Nuk shpjegon pse **mbijetojnë** — pse modelet, edhe pas gjithë trajnimit të mëvonshëm që synon t'i bëjë të dobishme dhe të ndershme, vazhdojnë të shtiren sikur e dinë përgjigjen në vend që të pranojnë pasigurinë. Këtu analogjia e punimit është pothuajse shqetësuese sa është e saktë. Mendo një nxënës në provim që nuk e di përgjigjen. Nëse një boshllëk merr zero pikë dhe një hamendësim mund të marrë një pikë, strategjia që maksimizon notën është të hamendësosh — me vetëbesim dhe hollësi, jo "nuk jam i sigurt". Nxënësit e mësojnë këtë. Edhe modelet, sepse i vlerësojmë njësoj. Autorët shqyrtuan benchmark-et ku fusha konkurron dhe leaderboard-et ku modelet optimizohen për të ngjitur, dhe gjetën se pothuajse të gjitha i japin "nuk e di" pikërisht të njëjtin rezultat si një përgjigje të gabuar: zero. Nën këtë rregull, një model që gjithmonë hamendëson do të mundë një model tjetër identik që sinjalizon ndershmërisht pasigurinë. Në një kuptim mjaft literal, **po i shpërblejmë modelet për të hamendësuar**.

Two ways to grade an answer

what each scoring rule rewards

Closed rubric

how most benchmarks score today

Correct	+1
Wrong	0
"I don't know"	0

Wrong and "I don't know" tie at 0, so a guess can only help.

Open rubric

scoring stated inside the question

Correct	+1
Wrong	−1
"I don't know"	0

A wrong guess now costs more than an honest "I don't know."

Benchmark-et mund ta bëjnë hamendësimin racional. Me skemën e zakonshme të pikëzimit (majtas), përgjigjja e gabuar dhe një "nuk e di" i ndershëm marrin të dyja zero — kështu një hamendësim vetëm mund të ndihmojë. Nëse rregullat shkruhen në vetë pyetjen, me penalizim për gabimin (djathtas), abstinimi kur modeli është i pasigurt mund të bëhet zgjedhja më e mirë. Kjo ndryshon atë që testi shpërblen; nuk e zgjidh vetvetiu problemin e halucinacioneve.

Original diagram — The Clean Paper CC BY 4.0

Kjo është pjesa që ia vlen të mbahet, sepse shkon kundër titullit të zakonshëm. Halucinacioni shpesh shitet si kufi i pashmangshëm dhe pothuajse misterioz i teknologjisë. Punimi e kundërshton këtë në të dy drejtimet. Dyshemeja e gabimit në pretraining nuk është mister — është gabim statistikor i zakonshëm, lloj problemi që machine learning e njeh prej dekadash. Dhe vazhdimësia e halucinacioneve nuk është e pashmangshme — është pjesërisht një stimul që **ne** e kemi ndërtuar dhe mund ta ndryshojmë. Një sistem që thjesht do të refuzonte të përgjigjej kur është i pasigurt nuk do të prodhonte përgjigje të rreme me vetëbesim në ato raste; arsyeja pse modelet e vendosura nuk sillen kështu është se scoreboard-et tona e penalizojnë refuzimin.

Çfarë bënë autorët

Punimi ka tri pjesë. Së pari, një argument matematik se një pjesë e halucinacionit është statistikisht e detyruar gjatë pretraining-ut, duke treguar se "gjenero vetëm tekst të vlefshëm" është të paktën aq i vështirë sa një problem binar klasifikimi "a është kjo deklaratë e vlefshme?". Së dyti, një argument — i mbështetur nga një shqyrtim i dhjetë benchmark-eve me ndikim — se metrikat kryesore të tipit "accuracy" shpërblejnë hamendësimin mbi abstinimin. Së treti, një zgjidhje e propozuar dhe një case study për ta testuar: **open rubrics**, ku mënyra e pikëzimit shkruhet brenda vetë pyetjes (për shembull, "një përgjigje e saktë merr 1, një e gabuar −1, prandaj abstino nëse je më pak se 50% i sigurt"), në mënyrë që modeli të dijë kur ndershmëria shpërblehet. Autorët e provojnë këtë me katër modele frontier — Gemini 3 Pro të Google, GPT-5 të OpenAI, Grok 4 të xAI dhe Claude Opus 4.5 të Anthropic — duke përdorur 4,326 pyetje faktike të SimpleQA. Ata e thonë qartë se case study është ilustrues, **jo vlerësim i kontrolluar ndërmjet modeleve** (settings default, pa tuning dhe pa normalizim të kostos).

Çfarë gjetën

- **Pretraining-u detyron njëfarë gabimi.** Norma me të cilën modeli prodhon pohime të rreme me vetëbesim ka një kufi të poshtëm të lidhur me gabimin e klasifikuesit më të mirë "a është kjo deklaratë e vlefshme?" që mund të ndërtohet prej tij. Për fakte pa model të mësueshëm, ky kufi është të paktën norma e *singleton*-ëve — pjesa e fakteve që shfaqen vetëm një herë në trajnim. Njëfarë gabimi është i pashmangshëm edhe me të dhëna të përsosura.
- **Pikëzimi shpërblen hamendësimin — në mënyrë konkrete.** Me pikëzim të zakonshëm saktë/gabim, mosabstinimi është strategji optimale dhe shqyrtimi i autorëve gjen se shumica dërrmuese e benchmark-eve të njohura e trajtojnë "nuk e di" thjesht si gabim. Një shembull i fortë nga modelet e vetë OpenAI-t: në SimpleQA, accuracy bruto favorizon pak o4-mini — që përgjigjet pothuajse gjithmonë dhe gabon në më shumë se tre të katërtat e rasteve — ndaj GPT-5-mini, që bën shumë më pak gabime sepse abstinon kur është i pasigurt. Modeli më i pakujdesshëm duket më mirë në scoreboard.
- **Open rubrics e përmbysin stimulimin në case study.** Autorët testojnë një teknikë të thjeshtë për uljen e halucinacioneve (modeli përgjigjet dy herë dhe abstinon nëse dy përgjigjet nuk pajto-

hen). Me accuracy standard, teknika ul gabimet **por ul edhe accuracy** — pra metrika e dekurajon përdorimin. Me open rubrics, e njëjta teknikë del më mirë për të katër modelet në një varg penalitetesh; dhe GPT-5-mini — që accuracy bruto e kishte penalizuar për abstinim — del më mirë se o4-mini kur rregullat e pikëzimit deklarohen hapur (n = 4,326 pyetje për model).

Çfarë ka të ngjarë të thotë kjo

Ulja e halucinacioneve nuk është kryesisht çështje për të shpikur gjithnjë e më shumë teste specifike për halucinacione. Është çështje për të ndryshuar si benchmark-et kryesore e vlerësojnë pasigurinë, që pranimi "nuk e di" të mos penalizohet më. Derisa scoreboard-i të ndryshojë, ulja e halucinacioneve do të vazhdojë t'u kushtojë modeleve pikë accuracy dhe, si pasojë, të dekurajohet — prandaj autorët e quajnë problemin "socio-teknik": pjesërisht duhet metrikë më e mirë, pjesërisht duhet që leaderboard-et me ndikim ta përdorin.

Çfarë *nuk* provon kjo

- **Nuk tregon se open rubrics i zgjidhin halucinacionet në përdorim real.** Eksperimenti mbështetës është një case study i vogël dhe qëllimisht **i pakontrolluar** — katër modele me settings default, një teknikë uljeje, një test faktik QA — i projektuar për të demonstruar **ndryshimin e stimulit**, jo për të renditur modelet ose për të provuar efektivitet të përgjithshëm.
- **Nuk pretendon se pikëzimi është shkaku i vetëm.** Gabimet në të dhënat e trajnimit, probleme vërtet të vështira dhe prompt-e të panjohura mbeten burime të tjera.
- **Nuk mbështet idenë popullore se halucinacionet janë të pashmangshme.** Autorët argumentojnë të kundërtën: një sistem që do t'u përgjigjej vetëm pyetjeve që mund t'i verifikojë dhe përndryshe do të thoshte "nuk e di" nuk do të halucinonte në ato raste.
- **Nuk e zhduk dyshemenë e gabimit të pretraining-ut** — e shpjegon dhe e kufizon, dhe kufiri vlen për gabime faktike me vetëbesim, jo për gjithë sjelljen e modelit.
- **Nuk tregon se open rubrics janë të mjaftueshme vetë.** Ato ndryshojnë atë që një vlerësim shpërblen; nuk zëvendësojnë retrieval, përdorimin e mjeteve ose kalibrimin më të mirë të modeleve.

Sa e fortë është prova?

- Bërthama është **matematikë** — kufij të poshtëm formalë, jo matje. Si argument teorik qëndron në kushtet e veta.
- Mbështetet në **modele qëllimisht të thjeshtuara** të problemit; autorët vetë sinjalizojnë "trikotominë e rreme" ku çdo përgjigje trajtohet si saktë, gabim ose "nuk e di", dhe mjedisin idealizuar të "fakteve arbitrare" që përdoret për kufirin më të pastër.
- Shqyrtimi i benchmark-eve është një **mostër e vogël dhe e kuruar** — dhjetë vlerësime me ndikim, jo auditim i plotë.

- Case study është **real por i kufizuar**: katër modele frontier, një teknikë uljeje, vetëm SimpleQA, settings default, shprehimisht "jo vlerësim i kontrolluar". Është proof of concept për argumentin e stimulit, jo rezultat benchmark-u.
- Ia vlen të emërtohet këndvështrimi: tre nga katër autorët janë ose kanë qenë të punësuar te **OpenAI** dhe punimi argumenton se fusha duhet të ndryshojë mënyrën si vlerëson modelet. Është një pozicion i argumentuar mirë nga një palë e interesuar, jo rishikim i jashtëm neutral — duhet peshuar, jo hedhur poshtë. (Në favor të tij, punimi e drejton kritikën po aq lehtë te modelet e veta, o4-mini dhe GPT-5-mini, sa te të tjerët.)

Pse ka rëndësi

Punimi e riformulon një problem shumë të ekzagjeruar. "Halucinacioni" shpesh shitet ose si defekt i frikshëm, ose si mur i pakapërcyeshëm; këtu bëhet diçka më e zakonshme dhe pjesërisht e krijuar prej nesh — një dysheme statistikore që mund ta kuptojmë, mbi të cilën vendoset një stimul që vetë e kemi zgjedhur. Mësimi më i gjerë është më i qetë dhe më i dobishëm: përparimi i mëtejshëm në besueshmëri mund të varet po aq nga **çfarë masim** sa nga çfarë ndërtojmë.

Përmbledhje e shkurtër

Përgjigjet e rreme me vetëbesim nga modelet gjuhësore vijnë nga dy burime. I pari është statistikor: kur një fakt nuk ka model për t'u mësuar, një model i trajnuar të imitojë gjuhën do të gabojë ndonjëherë dhe dyshemeja e gabimit mund të vlerësohet, për shembull nga sa fakte shfaqen vetëm një herë në trajnim. I dyti është stimulimi: pothuajse çdo benchmark ku modelet renditen e vlerëson "nuk e di" njësoj si një përgjigje të gabuar, kështu që hamendësimi gjithmonë ka më shumë shans të fitojë — deri në pikën ku një model që gabon në tre të katërtat e rasteve mund të kalojë një model më të ndershëm që abstinon. Propozimi i autorëve nuk është një test tjetër halucinacionesh, por **open rubrics**: deklaro mënyrën e pikëzimit brenda pyetjes. Në një case study me katër modele frontier, kjo e përmbys stimulimin dhe bën që një teknikë që ul halucinacionet të shpërblehet në vend që të penalizohet. Është punim teori-plus-shqyrtim me një eksperiment të vogël, shprehimisht të pakontrolluar, i peer-reviewed në *Nature*; zgjidhja është premtuese por ende nuk është demonstruar në shkallë, dhe halucinacionet argumentohen si as misterioze, as domosdoshmërisht të pashmangshme.

Kontroll pa teprime

Çfarë tregon punimi: Një kufi të poshtëm matematik nën të cilin njëfarë gabimi faktik gjatë pretraining-ut është i detyruar (të paktën "singleton rate" për fakte pa model); një shqyrtim që gjen se shumica e benchmark-eve kryesore nuk i japin kredi "nuk e di"; dhe një case study me katër modele ku deklarimi i pikëzimit në prompt ("open rubrics") bën që një metodë për uljen e halucinacioneve të fitojë, aty ku accuracy e thjeshtë e kishte penalizuar.

Çfarë është e besueshme, por jo e provuar: Që open rubrics, nëse integrohen në benchmark-et kryesore, do të ulin në mënyrë domethënëse halucinacionet në modelet e vendosura. Eksperimenti mbështetës është i vogël dhe shprehimisht i pakontrolluar.

Çfarë nuk tregon: Që halucinacionet janë të pashmangshme (argumenton të kundërtën); që pikëzimi është shkaku i vetëm; që halucinacionet mund të eliminohen plotësisht me këtë metodë; që case study rendit katër modelet kundër njëri-tjetrit.

Kufizimet kryesore: Model qëllimisht i thjeshtuar saktë/gabim/"nuk e di" (autorët e quajnë "trikotomi të rreme"); mostër e vogël benchmark-esh (dhjetë); case study i pakontrolluar (katër modele, një teknikë, një test, settings default); dhe një argument i udhëhequr nga autorë të lidhur me OpenAI për mënyrën si fusha duhet të vlerësojë modelet.

Sa besim duhet të ketë një lexues i përgjithshëm? Të lartë se halucinacioni nuk është as misterioz, as rreptësisht i pashmangshëm, dhe se benchmark-et kryesore aktualisht shpërblejnë hamendësimin. Mesatar se zgjidhja e propozuar ndihmon — tani ka proof of concept real, por jo ende demonstrim të gjerë dhe në shkallë.

Burimet

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