The Clean Paper

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

Zakaj jezikovni modeli halucinirajo — in zakaj jih način ocenjevanja pri tem ohranja

Samozavestne neresnice, ki jim pravimo »halucinacije«, niso skrivnostna napaka: nekatere so statističen stranski produkt učenja, vztrajajo pa tudi zato, ker običajni benchmarki nagradijo samozavesten ugibanje bolj kot pošten »ne vem«. Študija primera na štirih vodilnih modelih kaže, da zapis pravil točkovanja neposredno v promptu (»odprte rubrike«) ta spodbujevalni učinek obrne.

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

Zakaj modeli ugibajo — in kako smo jih tega naučili

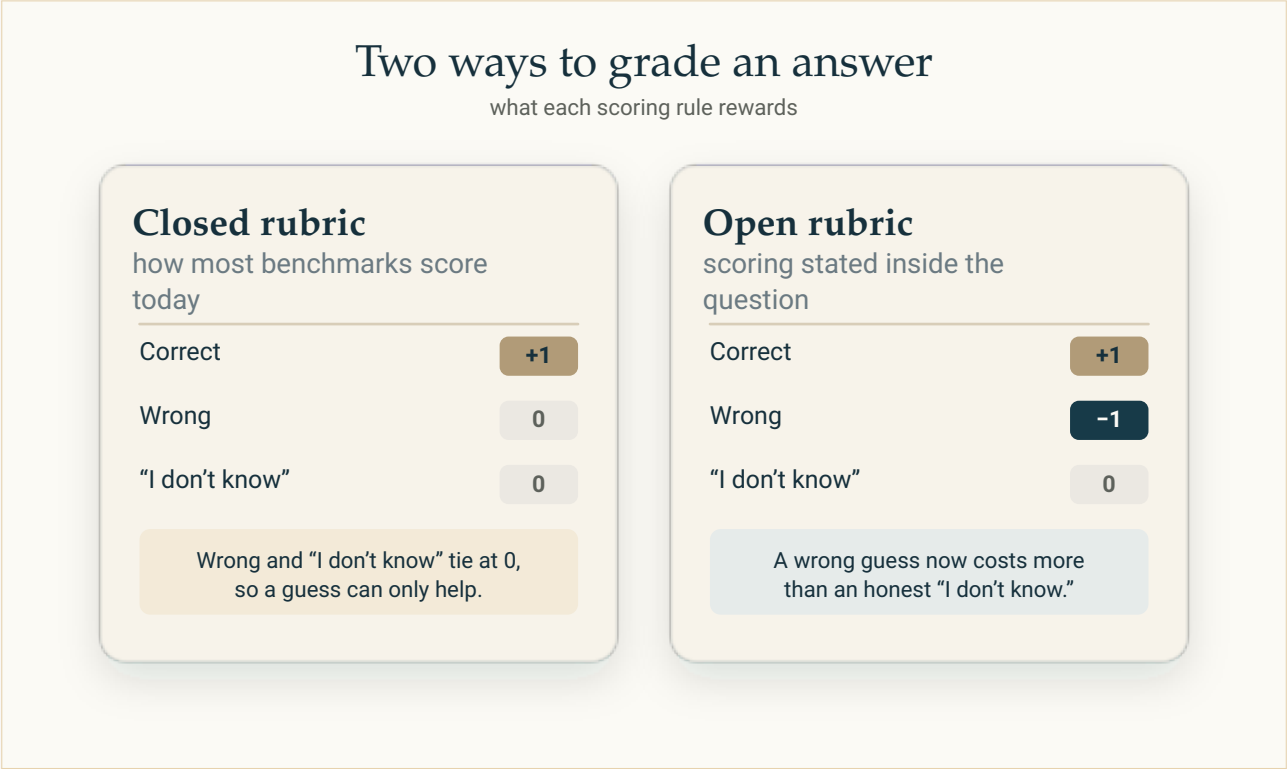
Vprašajte velik jezikovni model po rojstnem dnevu neznane osebe in lahko samozavestno odgovori »7. marec«, kot da datum bere s kartice — ter se zmoti, trikrat zapored, vsakič z drugim datumom. Avtorji navedejo prav tak primer: vodilni modeli dobijo preprosto faktografsko vprašanje — rojstni dan osebe ali pomen nejasne kratice — in vsak samozavestno izmisli drugačen odgovor, noben pravilen. Industrija temu pravi *halucinacija*, kar zveni kot napaka zaznave. Prva poteza članka je, da odstrani skrivnostnost.

Začnimo z učenjem modela. V prvi in največji fazi se v bistvu uči, kako izgleda tekoč jezik, tako da prebere ogromno besedila. Nato vzemimo dejstvo, za katerim ni vzorca — rojstni dan določene osebe. Če se je datum v učnem besedilu pojavil enkrat ali nikoli, se model, ki išče vzorce, nima česa oprijeti: z njegovega vidika je odgovor arbitraren. Avtorji to formalizirajo z izposojo stare ideje Alana Turinga iz drugega problema: če se eden od petih rojstnih dni v podatkih pojavi le enkrat, lahko pričakujemo, da se bo model zmotil pri najmanj enem od petih — ne zato, ker je pokvarjen, ampak ker ni bilo dovolj za učenje. (Po isti logiki modeli skoraj nikoli ne zgrešijo prestolnice države: takšni podatki se pojavljajo znova in znova.) Avtorji previdno pokažejo, da je že razlikovanje resnične izjave od verjetne lažne težaven problem in da je ustvarjanje *samo* resničnih izjav najmanj tako težko. Nek spodnji prag napak je vgrajen.

Ideja v ozadju: kako prešteti, česar še niste videli

V ozadju je res elegantna ideja — starejša od jezikovnih modelov in vredna lastne razlage.
Začnite z vrečo barvnih kroglic. Ne veste, koliko barv vsebuje. Sto kroglic izvlečete eno za drugo in jih preštejete: rdečih 40, modrih 25, zelenih 15, rumenih 5, vijoličnih 3, oranžnih 2 — nato pa **deset različnih barv, ki se vsaka pojavi samo enkrat**.
Vprašanje, s katerim se je Turing dejansko ukvarjal v povsem drugem kontekstu, je bilo: *kakšna je verjetnost, da bo naslednja kroglica barve, ki je še sploh niste videli?* Nevidenih barv ne morete prešteti — lahko pa preštejete barve, ki ste jih videli samo enkrat, tako imenovane **singletons**.
Pri **Good–Turingovi oceni** delež enkrat opaženih dogodkov ocenjuje verjetnostno maso, ki se še skriva med neopaženimi. Če je bilo deset od sto izvlečenih kroglic barv, ki so se pojavile le enkrat, je verjetnost, da bo naslednja barva povsem nova, približno **10 / 100 = 10 %**.
Te enkrat opažene barve niso *napake*. So merilo lastne nevednosti: če veliko barv vidite le enkrat, vam vzorec sporoča, da obstaja še veliko takih, ki jih sploh niste videli.
Zdaj barve zamenjajte z rojstnimi dnevi, vrečo pa z učnim besedilom modela. Če se med opaženimi rojstnimi dnevi **eden od petih pojavi samo enkrat**, enaka logika pravi, da približno petina verjetnosti pripada rojstnim dnevom, ki jih model praktično ni videl — pri rojstnem dnevu pa ni vzorca, iz katerega bi ga lahko izračunal. Datum, ki se pojavi enkrat ali nikoli, je zato dejstvo, ki ga model iz podatkov ne more zanesljivo rekonstruirati, in bo pri približno **enem od petih** zgrešil. Nobena dodatna »pamet« ne odpravi informacije, ki je v podatkih ni bilo.
To je argument v malem: delež singletonov meri, koliko sveta je *iz teh podatkov* neizučljivega, in postane **spodnja meja** napak. Prav zato model skoraj nikoli ne zgreši prestolnice: *Pariz* se pojavlja neštetokrat, njegov singleton rate je praktično nič in materiala za učenje je ogromno.

To pojasni izvor nekaterih halucinacij. Ne pojasni, zakaj *vztrajajo* — zakaj modeli tudi po poznejšem učenju, namenjenem koristnosti in poštenosti, še vedno blefirajo namesto da priznajo dvom. Tu je analogija članka neprijetno dobra. Predstavljajte si učenca na izpitu, ki odgovora ne pozna. Če prazen odgovor prinese nič točk, ugibanje pa lahko prinese eno, je za oceno optimalno ugibati — samozavestno, specifično, nikoli »nisem prepričan«. Učenci se tega naučijo. Izkazalo se je, da tudi modeli — ker jih ocenjujemo enako. Avtorji so pregledali benchmarke, na katerih področje dejansko tekmuje in na katerih se modeli optimizirajo, ter ugotovili, da skoraj vsi odgovor »ne vem« ocenijo povsem enako kot napačen odgovor: z ničlo. Po takem pravilu bo model, ki vedno ugiba, prehitel sicer enak model, ki pošteno izrazi negotovost. V precej dobesednem smislu jih z ocenjevanjem spodbujamo k ugibanju.



Benchmarki lahko naredijo ugibanje racionalno. Pri običajnem točkovanju (levo) napačen odgovor in pošten »ne vem« oba prineseta nič — zato ugibanje lahko samo pomaga. Če pravila napišemo v vprašanje in napačen odgovor kaznujemo (desno), lahko odstop ob negotovosti postane boljša strategija. To spremeni, kaj test nagrajuje; samo po sebi pa halucinacij ne reši.

Original diagram — The Clean Paper CC BY 4.0

Ta del je vredno obdržati, ker nasprotuje običajnemu naslovu. Halucinacije se pogosto predstavljajo kot neizogibna, skoraj mistična meja tehnologije. Članek nasprotuje obema deloma. Spodnja meja iz pretreniranja ni skrivnost — je običajna statistična napaka, kakršno strojno učenje razume že desetletja. Njihova vztrajnost pa ni neizogibna — deloma je posledica spodbude, ki smo jo sami zgradili in jo lahko spremenimo. Sistem, ki bi preprosto zavrnil odgovor, kadar ni prepričan, sploh ne bi haluciniral; razlog, da uvedeni modeli tega ne počnejo dosledno, je med drugim ta, da jih naše lestvice za zavrnitev kaznujejo.

Kaj so naredili avtorji

Članek ima tri dele. Prvi je matematični argument, da je del halucinacij statistično prisiljen že med preučenjem, saj pokaže, da je naloga »ustvari samo veljavno besedilo« najmanj tako težka kot binarna klasifikacija »ali je ta izjava veljavna?«. Drugi je argument — podprt s pregledom desetih vplivnih benchmarkov — da običajne metrike natančnosti nagradijo ugibanje bolj kot odstop. Tretji je predlagan popravek in študija primera: ocenjevanje z **odprtimi rubrikami**, kjer je pravilo točkovanja zapisano v samem vprašanju (na primer »pravilen odgovor = 1, napačen = −1, zato odstopi, če si manj kot 50 % prepričan«), tako da model ve, kdaj je poštenost nagrajena. To preizkusijo na štirih vodilnih modelih — Google Gemini 3 Pro, OpenAI GPT-5, xAI Grok 4 in Anthropic Claude Opus 4.5 — z 4.326 faktografskimi vprašanji SimpleQA. Izrecno napišejo, da je študija primera ilustrativna in »ni nadzorovana primerjava med modeli«: uporabljene so privzete nastavitve, brez prilagajanja in brez normalizacije stroškov.

Kaj so ugotovili

- **Preučenje prisili nekaj napak.** Stopnja, pri kateri model izda samozavestne neresnice, ima spodnjo mejo, povezano z (približno dvakratno) napako najboljšega klasifikatorja »ali je izjava veljavna?«, zgrajenega iz njega. Pri dejstvih brez učljivega vzorca je spodnja meja vsaj *singleton rate* — delež dejstev, ki se v učenju pojavijo samo enkrat. Del halucinacij je zato neizogiben tudi ob popolnoma čistih podatkih.
- **Ocenjevanje konkretno nagrajuje ugibanje.** Pri običajnem točkovanju pravilno/napačno je optimalno nikoli ne odstopiti, pregled avtorjev pa pokaže, da velika večina priljubljenih benchmarkov »ne vem« preprosto oceni kot napačno. Nazoren primer iz njihovih lastnih modelov: na SimpleQA surova natančnost rahlo *favorizira* OpenAI o4-mini — ki odgovori skoraj na vse in se moti v več kot treh četrtinah primerov — pred GPT-5-mini, ki naredi bistveno manj napak, ker ob negotovosti odstopi. Bolj brezglavi model je na lestvici videti boljši.
- **Odprte rubrike v študiji primera obrnejo spodbudo.** Avtorji preizkusijo preprosto blaženje halucinacij: model odgovori dvakrat in odstopi, če se odgovora ne ujemata. Pri standardni natančnosti metoda zmanjša napake, vendar zmanjša tudi rezultat natančnosti — metrika torej odvrača od njene uporabe. Pri odprtih rubrikah ista metoda prehiti osnovno različico pri vseh štirih modelih čez razpon kazni; GPT-5-mini, ki ga je surova natančnost kaznovala zaradi odstopanja ob negotovosti, pa prehiti o4-mini, ko je pravilo točkovanja zapisano odkrito (n = 4.326 vprašanj na model).

Kaj to verjetno pomeni

Zmanjševanje halucinacij verjetno ni predvsem problem izmišljanja vedno novih posebnih testov za halucinacije. Gre za spremembo načina, kako glavni benchmarki ocenjujejo negotovost, tako da priznanje »ne vem« ni več kaznovano. Dokler se lestvice ne spremenijo, bo zmanjševanje halucinacij

modelom še naprej jemalo točke natančnosti in bo zato sistemsko nezaželeno. Zato avtorji problem imenujejo »družbeno-tehničen«: del rešitve je boljša metrika, del pa prepričati vplivne lestvice, naj jo sprejmejo.

Česa to ne dokazuje

- **Ne** kaže, da odprte rubrike odpravijo halucinacije v resnični uporabi. Podporni poskus je majhna, namenoma **nenadzorovana** študija primera — štirje modeli s privzetimi nastavitvami, ena metoda blaženja in en faktografski test — namenjena prikazu *obrata spodbude*, ne rangiranju modelov ali dokazovanju splošne učinkovitosti.
- **Ne** trdi, da je ocenjevanje *edini* vzrok. Napake v učnih podatkih, resnično težki problemi in neznani prompti ostajajo ločeni viri.
- **Ne** podpira priljubljene trditve, da so halucinacije neizogibne. Avtorji trdijo nasprotno: sistem, ki bi odgovarjal le na preverljiva vprašanja in sicer rekel »ne vem«, ne bi haluciniral.
- **Ne** odpravi spodnje meje iz preučenja — pojasni jo in omeji; meja se nanaša na samozavestne faktografske napake, ne na vse vedenje modela.
- **Ne** kaže, da so odprte rubrike same po sebi *zadostne*. Spremenijo, kaj ocena nagrajuje; niso nadomestilo za iskanje podatkov, uporabo orodij ali bolje kalibrirane modele.

Kako močni so dokazi?

- Jedro je **matematika** — formalne spodnje meje, ne meritve. Kot teoretični argument je znotraj svojih predpostavk trden.
- Temelji na **namenoma poenostavljenih modelih** problema; avtorji sami opozorijo na »lažno trihotomijo«, ko vsak odgovor obravnavamo kot pravilen, napačen ali »ne vem«, ter na idealizirano okolje »arbitrarnih dejstev«, uporabljeno za najčistejšo mejo.
- Pregled benchmarkov je **majhen, kuriran vzorec** — deset vplivnih evalvacij, ne izčrpen popis.
- Študija primera je **resnična, a omejena**: štirje vodilni modeli, ena metoda blaženja, samo SimpleQA, privzete nastavitve in izrecno »ne nadzorovana evalvacija«. Gre za dokaz načela glede spodbud, ne benchmark rezultat.
- Vredno je navesti tudi perspektivo: trije od štirih avtorjev so ali so bili zaposleni pri **OpenAI**, članek pa zagovarja spremembo načina, kako področje ocenjuje modele. To je dobro argumentirano stališče zainteresirane strani, ne nevtralni zunanji pregled — nekaj, kar je treba tehtati, ne zavrniti. Članku v prid: kritiko enako hitro usmeri v lastna modela o4-mini in GPT-5-mini kot v druge.

Zakaj je pomembno

Močno preokviri pretirano izpostavljen problem. »Halucinacija« se rada prodaja ali kot strašljiv hrošč ali kot nepremagljiv zid; članek jo naredi navadno in deloma samo-povzročeno — statistično spodnjo

mejo, ki jo znamo razumeti, na vrhu pa spodbudo, ki smo jo izbrali sami. Širša lekcija je tišja in uporabnejša: nadaljnji napredek pri zanesljivosti je lahko prav toliko odvisen od *tega, kaj merimo*, kot od tega, kaj gradimo.

Kratek povzetek

Samozavestni napačni odgovori jezikovnih modelov prihajajo iz dveh smeri. Prva je statistična: ko dejstvo nima vzorca, ki bi ga bilo mogoče naučiti, ga bo model, naučen posnemati jezik, včasih zgrešil, spodnjo mejo pa lahko ocenimo, denimo po tem, koliko dejstev se v učnih podatkih pojavi samo enkrat. Druga so spodbude: skoraj vsak benchmark, po katerem modele rangiramo, oceni »ne vem« enako kot napačen odgovor, zato je ugibanje vedno boljša strategija — do te mere, da lahko model, ki se moti v treh četrtinah primerov, prehiti bolj pošten model, ki ob negotovosti odstopi. Predlog avtorjev ni še en test halucinacij, temveč »odprte rubrike«: pravilo točkovanja napišite v samo vprašanje. V študiji primera na štirih vodilnih modelih to obrne spodbudo, tako da je metoda, ki zmanjša halucinacije, nagrajena namesto kaznovana. Gre za teoretični članek s pregledom in majhnim, izrecno nenadzorovanim poskusom, recenziran v *Nature*; popravek je obetaven, vendar še ni dokazano učinkovit v velikem obsegu, halucinacije pa po argumentu niso niti skrivnostne niti strogo neizogibne.

Preverjanje brez olepševanja

Kaj članek pokaže: Matematično spodnjo mejo, zaradi katere je nekaj halucinacij med preučenjem statistično prisiljenih (pri dejstvih brez vzorca vsaj singleton rate); pregled, ki ugotovi, da večina vodilnih benchmarkov »ne vem« ne nagradi; in študijo primera štirih modelov, v kateri zapis točkovanja v promptu (»odprta rubrika«) omogoči, da metoda za zmanjšanje halucinacij zmaga tam, kjer jo je gola natančnost kaznovala.

Kaj je verjetno, vendar ne dokazano: Da bi odprte rubrike, vključene v glavne benchmarke, občutno zmanjšale halucinacije v uvedenih modelih. Podporni eksperiment je majhen in izrecno nenadzorovan.

Česa ne pokaže: Da so halucinacije neizogibne — trdi nasprotno; da je ocenjevanje edini vzrok; da je mogoče halucinacije povsem odpraviti; da študija primera rangira štiri modele med seboj.

Glavne omejitve: Namenoma poenostavljen model pravilen/napačen/»ne vem« (avtorji ga sami imenujejo »lažna trihotomija«); majhen kuriran pregled benchmarkov (deset evalvacij); nenadzorovana študija primera (štirje modeli, ena metoda, en test, privzete nastavitve); ter argument, ki ga vodijo avtorji iz OpenAI o tem, kako bi področje moralo ocenjevati modele.

Koliko zaupanja naj ima splošni bralec? Visoko, da halucinacije niso niti skrivnostne niti strogo neizogibne in da glavni benchmarki trenutno nagrajujejo ugibanje. Zmerno, da predlagani popravek pomaga — zdaj ima resničen dokaz načela, ne pa še dokaza široke učinkovitosti v velikem obsegu.

Viri

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