The Clean Paper

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

zašto jezički modeli haluciniraju — i zašto ih način na koji ih ocenjujemo u tome održava

Samouverene neistine koje nazivamo „halucinacijama" nisu misteriozni kvar: neke su statistički nusproizvod treniranja, a opstaju zato što glavni benčmarkovi nagrađuju samouvereno pogađanje više od poštenog „ne znam". Studija slučaja na četiri najnaprednija modela pokazuje da navođenje pravila bodovanja u promptu („otvorene šeme bodovanja") okreće taj podsticaj.

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

Zašto modeli pogađaju — i zašto smo ih tome naučili

Pitajte veliki jezički model za datum rođenja nepoznate osobe i možda će odgovoriti „7. ožujka" mirnom sigurnošću nekoga ko ga čita s kartice — i pogrešiti, tri puta zaredom, s tri različita datuma. Autori daju upravo takve primere: vodeći modeli dobiju jednostavno činjenično pitanje — nečiji rođendan ili značenje opskurne kratice — i svaki samouvereno izmisli drugačiji odgovor, a nijedan nije tačan. Industrija to naziva *halucinacijom,* što zvuči kao kvar percepcije. Prvi potez rada je ukloniti misterij.

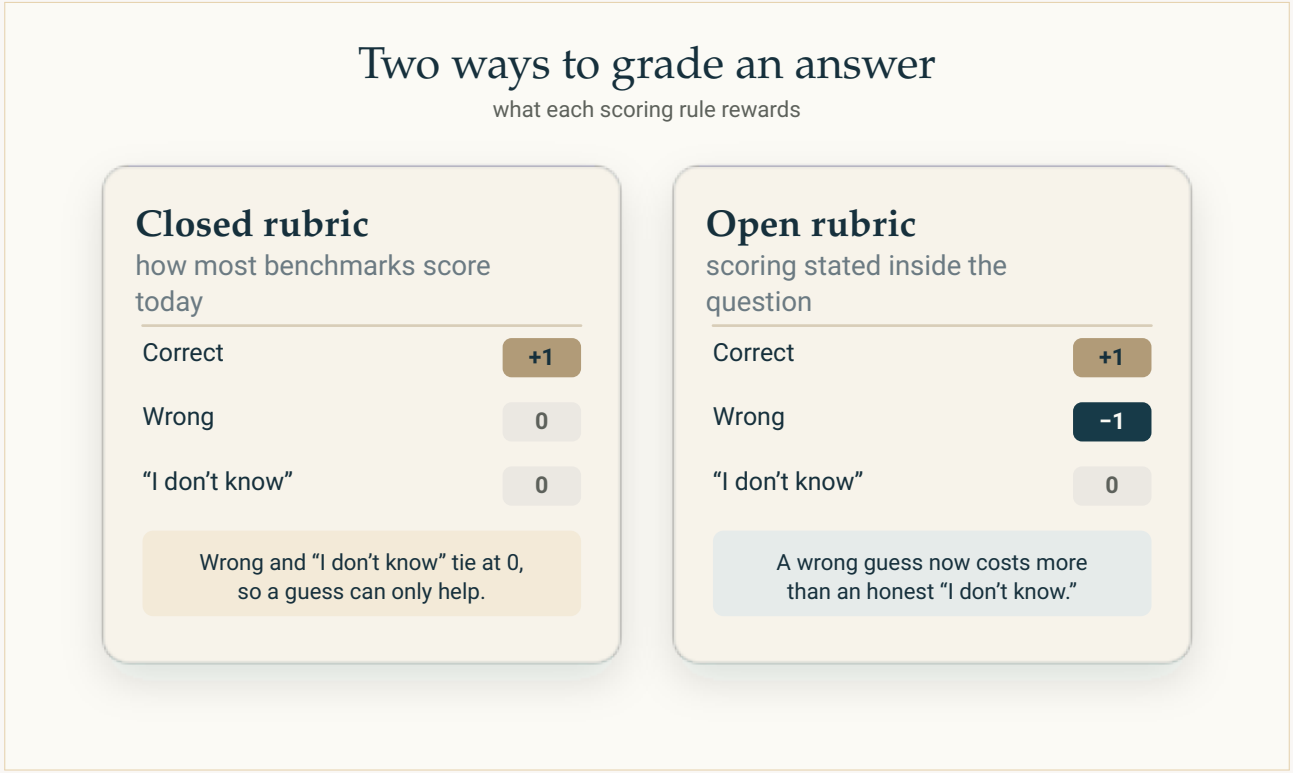
Počnimo od toga kako model nastaje. U prvoj i najvećoj fazi treniranja uči, pojednostavljeno, kako izgleda tečan jezik čitajući ogromnu količinu teksta. Sada uzmite činjenicu bez uzorka — rođendan jedne određene osobe. Ako se taj datum pojavio u podacima jednom ili nikada, model koji uči obrasce nema za što uhvatiti: s njegove je tačke gledišta odgovor proizvoljan. Autori to formaliziraju posuđujući staru ideju Alana Turinga iz drugog problema: ako se jedan od pet rođendana u podacima pojavi samo jednom, treba očekivati da će model pogrešiti barem jedan od pet — ne zato što je pokvaren, nego zato što nije bilo dovoljno materijala za učenje. (Po istoj logici modeli gotovo nikada ne greše glavni grad države: ti se podaci pojavljuju stalno.) Autori pažljivo argumentiraju da je razlikovati istinitu tvrdnju od uverljive lažne tvrdnje samo po sebi težak problem i da je proizvoditi *isključivo* istinite tvrdnje barem jednako teško. Postoji donja granica greške.

Ideja ispod svega: kako brojiti ono što još niste videli

Oslanja se na zaista domišljatu ideju — stariju od jezičkih modela i vrednu upoznavanja.
Počnite s vrećom obojenih kuglica. Ne znate koliko boja sadrži. Izvučete 100 kuglica, jednu po jednu, i prebrojite ih: crvena 40, plava 25, zelena 15, žuta 5, ljubičasta 3, narandžasta 2 — a zatim **deset različitih boja koje se svaka pojave tačno jednom.**
Pitanje s kojim se Turing zapravo susreo, na sasvim drugom problemu, bilo je: *kolika je verovatnoća da je sledeća kuglica boje koju još niste videli?* Ne možete prebrojiti ono što nikada niste izvukli — ali **možete** prebrojiti boje koje ste videli tačno jednom, „singletons". Trik, nazvan **Good–Turingova procena**, kaže da udeo vaših izvlačenja koja su se pojavila samo jednom procenjuje verovatnoću još skrivenu u bojama koje niste videli. Deset od sto izvlačenja bile su jednokratne boje, pa je šansa da sledeća kuglica bude potpuno nova boja oko **10 / 100 = 10%.**
Te jednom viđene boje nisu *greške*. One mere sopstveno neznanje: mnogo boja koje se pojave jednom način je na koji vam uzorak govori da svet sadrži još mnogo toga što jednostavno niste izvukli.
Sada zamenite boje rođendanima, a vreću tekstom za treniranje modela. Pretpostavimo da se među rođendanima koje je model vidio **jedan od pet pojavljuje tačno jednom.** Isti trik: otprilike petina verovatnoće leži u rođendanima koje model praktično nikada nije vidio — a rođendan nema uzorak na koji se možete osloniti (ne možete nečiji rođendan *izračunati*). Datum viđen jednom ili nikada je kovanica koju model ne zna ponderirati, pa će pogrešiti na otprilike **jednom od pet**. Nikakva domišljatost to ne popravlja: nije bilo ničega što bi se moglo naučiti.
To je celi argument u malom: stopa jednokratnih pojavljivanja meri koliko je sveta nemoguće

> naučiti *iz ovih podataka* i to postavlja **donju granicu** greške. Zato model gotovo nikada ne pogreši glavni grad — *Paris* se pojavljuje stalno, njegova stopa jednokratnog pojavljivanja gotovo je nula i ima obilje materijala za učenje.

To objašnjava odakle halucinacije dolaze. Ne objašnjava zašto *prežive* — zašto modeli, nakon svih kasnijih faza treniranja koje ih trebaju učiniti korisnima i poštenima, i dalje blefiraju umesto da priznaju sumnju. Ovde je analogija rada gotovo neprijatno pogođena. Zamislite učenika na ispitu koji ne zna odgovor. Ako prazno polje nosi nula bodova, a pogađanje možda donosi jedan, strategija koja maksimalizira ocenu je pogađati — samouvereno, konkretno, nikada „nisam siguran". Učenici to nauče. I modeli, ispada, takođe — jer ih ocenjujemo na isti način. Autori su pregledali benčmarke na kojima se polje stvarno takmiči, skale prema kojima se modeli optimiraju, i pronašli da gotovo svi daju „ne znam" potpuno isti rezultat kao pogrešnom odgovoru: nulu. Pod tim pravilom model koji uvek pogađa nadmašiće inače identičan model koji pošteno označava svoju nebezbednost. U prilično doslovnom smislu, bodujemo ih prema blefiranju.



Benčmarkovi mogu pogađanje učiniti racionalnim. U uobičajenom bodovanju (levo) pogrešan odgovor i pošteno „ne znam" nose nula bodova — pa pogađanje može samo pomoći. Ako se pravila navedu u samom pitanju i pogrešan odgovor nosi kaznu (desno), odustajanje kada model nije siguran može postati bolji potez. To menja ono što test nagrađuje; samo po sebi ne rešava halucinacije.

Ovo vredi zadržati jer ide protiv uobičajenog naslova. Halucinacija se često predstavlja kao neizbežna, gotovo mistična granica tehnologije. Rad osporava oba dela. Donja granica iz predtreniranja nije misterij — to je obična statistička greška, kakvu mašinsko učenje razume decenijama. A njeno preživl-

javanje nije neizbežno — delimično je podsticaj koji smo sami izgradili i mogli bismo ga promeniti. Sistem koji bi jednostavno odbio odgovor kada nije siguran ne bi halucinirao; razlog zbog kojeg se modeli u upotrebi tako ne ponašaju je to što naši rezultati kažnjavaju odbijanje.

Šta su autori uradili

Rad ima tri dela. Prvo, matematički argument da je određena količina halucinacija statistički prisiljena tokom predtreniranja: pokazuje da je „generiši samo valjan tekst" barem jednako teško kao binarna klasifikacija „da li je ova tvrdnja valjana?". Drugo, argument — potkrepljen pregledom deset uticajnih benčmarkova — da uobičajene metrike tačnosti nagrađuju pogađanje umesto odustajanja. Treće, predloženo rešenje i studija slučaja koja ga testira: **otvorene šeme bodovanja** (*open rubrics*), gde je bodovanje napisano u samom pitanju, primerice „tačan odgovor nosi 1, pogrešan −1, pa odustani ako si manje od 50% siguran", tako da model zna kada se poštenje nagrađuje. To testiraju na četiri najnaprednija modela — Googleovu Gemini 3 Pro, OpenAI-jevu GPT-5, xAI-jevu Grok 4 i Anthropicovu Claude Opus 4.5 — koristeći 4,326 činjeničnih pitanja iz SimpleQA. Izričito navode da je studija slučaja ilustrativna, „nije kontrolisana evaluacija modela" (podrazumevane postavke, bez podešavanja i bez normalizacije troška).

Šta su pronašli

- **Predtreniranje prisiljava deo grešaka.** Stopa kojom model proizvodi samouverene neistine odozdo je ograničena otprilike dvostrukom stopom greške najboljeg klasifikatora „da li je ova tvrdnja valjana?" izgrađenog iz njega. Za činjenice bez naučivog uzorka ta donja granica najmanje je *stopa jednokratnih pojavljivanja* — udeo činjenica koje se u treningu pojavljuju tačno jednom. Deo halucinacija neizbežan je čak i uz savršeno čiste podatke.
- **Bodovanje konkretno nagrađuje pogađanje.** Uz obično bodovanje tačno/pogrešno, nikada ne odustati optimalna je strategija, a pregled autora pokazuje da velika većina popularnih benčmarkova „ne znam" jednostavno boduje kao pogrešno. Upečatljiv primer iz njihova sopstvenog dvorišta: na SimpleQA testu sirova tačnost malo *favorizira* OpenAI-jev o4-mini — koji odgovara gotovo na sve i greši više od tri četvrtine vremena — u odnosu na GPT-5-mini, koji čini mnogo manje grešaka jer odustaje kada nije siguran. Neoprezniji model na skali izgleda bolji.
- **Otvorene šeme bodovanja okreću podsticaj u njihovoj studiji slučaja.** Testiraju jednostavnu tehniku smanjenja halucinacija: model odgovori dvaput i odustane ako se odgovori ne slažu. Uz standardnu tačnost tehnika smanjuje greške, ali smanjuje i tačnost — pa metrika obeshrabruje njenu upotrebu. Uz otvorene šeme bodovanja ista tehnika pobeđuje za sva četiri modela kroz raspon kazni; i GPT-5-mini, kojeg je sirova tačnost kažnjavala zbog odustajanja, završava ispred o4-mini kada je bodovanje otvoreno navedeno (n = 4,326 pitanja po modelu).

Šta ovo verovatno znači

Smanjenje halucinacija uglavnom nije pitanje izmišljanja još testova specifičnih za halucinacije. Reč je o promeni načina na koji glavni benčmarkovi boduju nebezbednost, tako da priznanje „ne znam" više nije kažnjeno. Dok se skala ne promeni, smanjivanje halucinacija i dalje će modelima oduzimati bodove tačnosti i time biti obeshrabrivano — zato autori problem nazivaju „sociotehničkim": delom je potrebna bolja metrika, a delom prihvatanje te metrike na uticajnim skalama.

Šta ovo *ne* dokazuje

- **Ne** pokazuje da otvorene šeme bodovanja rešavaju halucinacije u stvarnom svetu. Prateći eksperiment mala je, namerno **nekontrolisana** studija slučaja — četiri modela na zadanim postavkama, jedna tehnika ublažavanja, jedan test činjeničnih pitanja — zamišljena da pokaže *okret podsticaja*, a ne rangira modele ili dokaže opštu efikasnost.
- **Ne** tvrdi da je bodovanje *jedini* uzrok. Greške u podacima za treniranje, stvarno teški problemi i nepoznati promptovi ostaju zasebni izvori.
- **Ne** podržava popularnu tvrdnju da su halucinacije neizbežne. Autori tvrde suprotno: sistem koji bi odgovarao samo na proverljiva pitanja i inače govorio „ne znam" nikada ne bi halucinirao.
- **Ne** uklanja donju granicu predtreniranja — objašnjava je i omeđuje, a granica se odnosi na samouverene činjenične greške, ne na sve ponašanje modela.
- **Ne** pokazuje da su otvorene šeme bodovanja same po sebi *dovoljne*. Menjaju ono što evaluacija nagrađuje; nisu zamena za dohvat podataka, alate ili bolje kalibrirane modele.

Koliko su dokazi jaki?

- Jezgra je **matematika** — formalne donje granice, ne merenja. Kao teorijski argument stoji na sopstvenim pretpostavkama.
- Oslanja se na **namerno pojednostavljene modele** problema; autori sami upozoravaju na „lažnu trihotomiju" u kojoj je svaki odgovor tačan, pogrešan ili „ne znam", i na idealizovano okruženje „proizvoljnih činjenica" korišćeno za najčišću granicu.
- Pregled benčmarkova je **mali, kurirani uzorak** — deset uticajnih evaluacija, ne iscrpan audit.
- Studija slučaja je **stvarna, ali ograničena**: četiri najnaprednija modela, jedna tehnika ublažavanja, samo SimpleQA, podrazumevane postavke, izričito „nije kontrolisana evaluacija". To je dokaz koncepta za argument o podsticajima, ne benčmark rezultat.
- Vredi navesti perspektivu autora: trojica od četvorice autora jesu ili su bili zaposlenici **OpenAI-ja**, a rad tvrdi da bi polje trebalo promeniti način evaluacije modela. To je dobro argumentirana pozicija zainteresirane strane, ne neutralna spoljašnja revizija — treba je odvagnuti, ne odbaciti. (U prilog radu, kritiku jednako usmerava prema sopstvenim modelima o4-mini i GPT-5-mini kao i prema drugima.)

Zašto je to važno

Rad preoblikuje jako reklamiran problem. „Halucinacija" se obično prodaje ili kao sablasni kvar ili kao nepomičan zid; ovaj rad je čini običnom i delimično samonametnutom — statistička donja granica koju možemo razumeti, nadograđena podsticajem koji smo sami odabrali. Šira pouka tiša je i korisnija: dalji napredak u pouzdanosti mogao bi zavisiti jednako o tome *što merimo* kao i o tome što gradimo.

Kratak rezime

Samouvereni lažni odgovori jezičkih modela dolaze iz dva izvora. Prvi je statistički: kada činjenica nema uzorak koji se može naučiti, model treniran na oponašanju jezika ponekad će pogrešiti, a ta se donja granica može proceniti, primerice iz broja činjenica koje se u treningu pojavljuju samo jednom. Drugi su podsticaji: gotovo svaki benčmark prema kojem se modeli rangiraju boduje „ne znam" isto kao pogrešan odgovor, pa pogađanje uvek pobeđuje — toliko da model koji greši tri četvrtine vremena može nadmašiti pošteniji model koji odustaje. predlog autora nije još jedan test halucinacija nego „otvorene šeme bodovanja": navesti bodovanje u samom pitanju. U studiji slučaja na četiri najnaprednija modela to okreće podsticaj tako da se metoda za smanjenje halucinacija nagrađuje umesto kažnjava. Reč je o teorijskom radu s pregledom i malim, izričito nekontrolisanim eksperimentom, recenziranom u časopisu *Nature*; rešenje je obećavajuće, ali još nije pokazano u velikom opsegu, a halucinacije prema argumentu nisu ni misteriozne ni strogo neizbežne.

Provera bez ulepšavanja

Šta rad pokazuje: Matematičku donju granicu prema kojoj je deo halucinacija prisiljen tokom predtreniranja, najmanje „stopu jednokratnih pojavljivanja" za činjenice bez uzorka; pregled koji nalazi da većina vodećih benčmarkova ne daje nikakav bod za „ne znam"; i studiju slučaja s četiri modela u kojoj navođenje bodovanja u promptu, „otvorene šeme bodovanja", omogućuje metodi za smanjenje halucinacija da pobedi tamo gde ju je obična tačnost kažnjavala.

Šta je verovatno, ali nije dokazano: Da bi otvorene šeme bodovanja, ako se ugrade u glavne benčmarkove, značajno smanjile halucinacije u modelima u upotrebi. Prateći eksperiment mali je i izričito nekontrolisan.

Šta ne pokazuje: Da su halucinacije neizbežne (tvrdi suprotno); da je bodovanje jedini uzrok; da se halucinacije mogu potpuno ukloniti; da studija slučaja rangira četiri modela jedan protiv drugoga.

Glavna ograničenja: Namerno pojednostavljen model tačno/pogrešno/„ne znam", koji autori nazivaju „lažnom trihotomijom"; mali kurirani pregled benčmarkova (deset evaluacija); nekontrolisana studija slučaja (četiri modela, jedna tehnika, jedan test, podrazumevane postavke); i argument predvođen OpenAI-jevim autorima o tome kako bi polje trebalo evaluirati modele.

Koliko poverenja treba imati opšti čitalac? Visoko da halucinacije nisu ni misteriozne ni strogo neizbežne i da glavni benčmarkovi trenutno nagrađuju pogađanje. Umereno da predloženo rešenje pomaže — sada ima stvaran dokaz koncepta, ali još ne i demonstraciju široke efikasnosti u velikom opsegu.

Izvori

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