The Clean Paper

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

Zašto jezični modeli haluciniraju — i zašto ih način na koji ih ocjenjujemo u tome održava

Samouvjerene neistine koje nazivamo „halucinacijama" nisu misteriozni kvar: neke su statistički nusproizvod treniranja, a opstaju zato što glavni benchmarkovi nagrađuju samouvjereno pogađanje više od poštenog „ne znam". Studija slučaja na četiri frontier modela pokazuje da navođenje pravila bodovanja u promptu („otvorene rubrike") okreće taj poticaj.

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

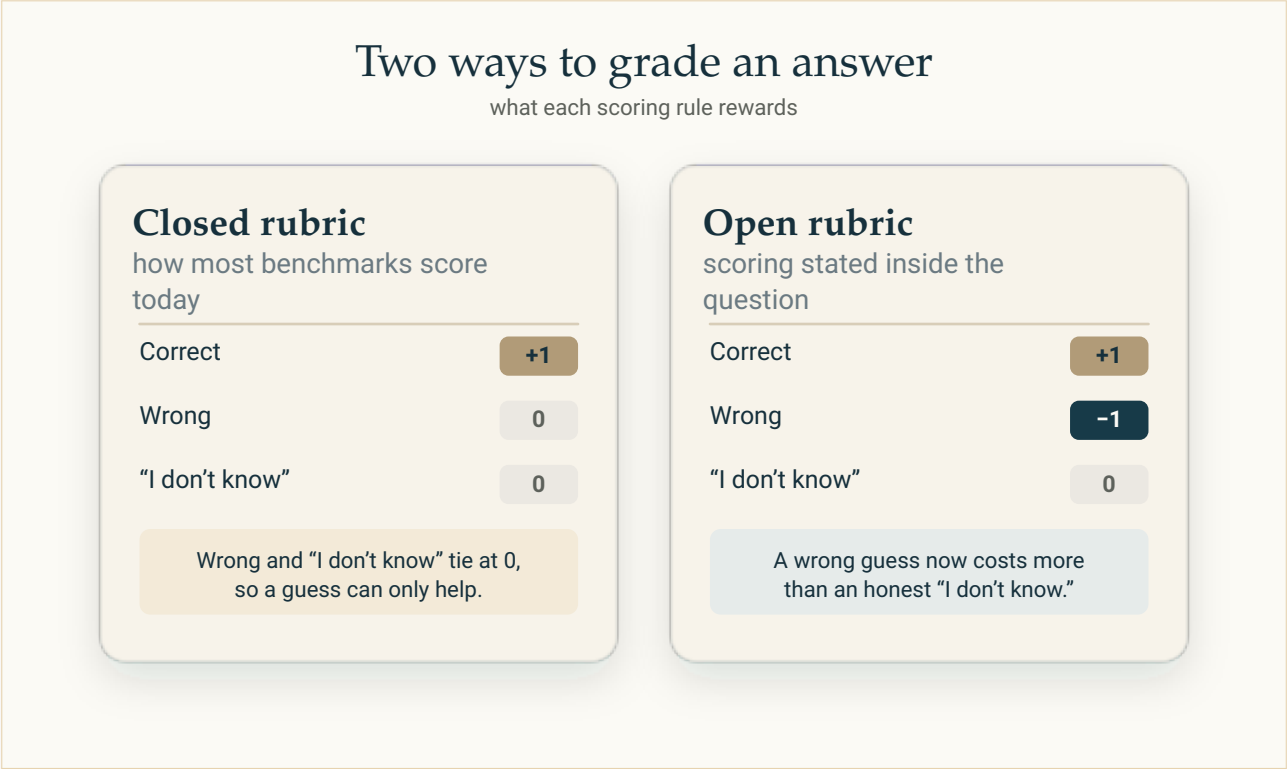
Počnimo od toga kako model nastaje. U prvoj i najvećoj fazi treniranja uči, pojednostavljeno, kako izgleda tečan jezik čitajući ogromnu količinu teksta. Sada uzmite činjenicu bez uzorka — rođendan jedne određene osobe. Ako se taj datum pojavio u podacima jednom ili nikada, model koji uči obrasce nema za što uhvatiti: s njegove je točke 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 pogriješiti 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 griješe glavni grad države: ti se podaci pojavljuju stalno.) Autori pažljivo argumentiraju da je razlikovati istinitu tvrdnju od uvjerljive lažne tvrdnje samo po sebi težak problem i da je proizvoditi *isključivo* istinite tvrdnje barem jednako teško. Postoji donja granica pogreške.

Ideja ispod svega: kako brojiti ono što još niste vidjeli

Oslanja se na doista domišljatu ideju — stariju od jezičnih modela i vrijednu 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, narančasta 2 — a zatim **deset različitih boja koje se svaka pojave točno jednom.**
Pitanje s kojim se Turing zapravo susreo, na sasvim drugom problemu, bilo je: *kolika je vjerojatnost da je sljedeća kuglica boje koju još niste vidjeli?* Ne možete prebrojiti ono što nikada niste izvukli — ali **možete** prebrojiti boje koje ste vidjeli točno jednom, „singletons". Trik, nazvan **Good–Turingova procjena**, kaže da udio vaših izvlačenja koja su se pojavila samo jednom procjenjuje vjerojatnost još skrivenu u bojama koje niste vidjeli. Deset od sto izvlačenja bile su jednokratne boje, pa je šansa da sljedeća kuglica bude potpuno nova boja oko **10 / 100 = 10%.**
Te jednom viđene boje nisu *pogreške*. One mjere vlastito neznanje: mnogo boja koje se pojave jednom način je na koji vam uzorak govori da svijet sadrži još mnogo toga što jednostavno niste izvukli.
Sada zamijenite boje rođendanima, a vreću tekstom za treniranje modela. Pretpostavimo da se među rođendanima koje je model vidio **jedan od pet pojavljuje točno jednom.** Isti trik: otprilike petina vjerojatnosti leži u rođendanima koje model praktički 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 pogriješiti na otprilike **jednom od pet**. Nikakva domišljatost to ne popravlja: nije bilo ničega što bi se moglo naučiti.
To je cijeli argument u malom: stopa jednokratnih pojavljivanja mjeri koliko je svijeta

> nemoguće naučiti *iz ovih podataka* i to postavlja **donju granicu** pogreške. Zato model gotovo nikada ne pogriješi 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 umjesto da priznaju sumnju. Ovdje je analogija rada gotovo neugodno pogođena. Zamislite učenika na ispitu koji ne zna odgovor. Ako prazno polje nosi nula bodova, a pogađanje možda donosi jedan, strategija koja maksimalizira ocjenu jest pogađati — samouvjereno, konkretno, nikada „nisam siguran". Učenici to nauče. I modeli, ispada, također — jer ih ocjenjujemo na isti način. Autori su pregledali benchmarke na kojima se polje stvarno natječe, ljestvice 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 uvijek pogađa nadmašit će inače identičan model koji pošteno označava svoju nesigurnost. U prilično doslovnom smislu, bodujemo ih prema blefiranju.



Benchmarkovi mogu pogađanje učiniti racionalnim. U uobičajenom bodovanju (lijevo) 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 mijenja ono što test nagrađuje; samo po sebi ne rješava halucinacije.

Original diagram — The Clean Paper CC BY 4.0

Ovo vrijedi zadržati jer ide protiv uobičajenog naslova. Halucinacija se često predstavlja kao neizbježna, gotovo mistična granica tehnologije. Rad osporava oba dijela. Donja granica iz predtreniranja nije misterij — to je obična statistička pogreška, kakvu strojno učenje razumije desetljećima. A

njezino preživljavanje nije neizbježno — djelomično je poticaj koji smo sami izgradili i mogli bismo ga promijeniti. Sustav koji bi jednostavno odbio odgovor kada nije siguran ne bi halucinirao; razlog zbog kojeg se modeli u uporabi tako ne ponašaju jest to što naši rezultati kažnjavaju odbijanje.

Što su autori napravili

Rad ima tri dijela. Prvo, matematički argument da je određena količina halucinacija statistički prisiljena tijekom predtreniranja: pokazuje da je „generiraj samo valjan tekst" barem jednako teško kao binarna klasifikacija „je li ova tvrdnja valjana?". Drugo, argument — potkrijepljen pregledom deset utjecajnih benchmarkova — da uobičajene metrike točnosti nagrađuju pogađanje umjesto odustajanja. Treće, predloženo rješenje i studija slučaja koja ga testira: **otvorene rubrike** (*open rubrics*), gdje je bodovanje napisano u samom pitanju, primjerice „točan 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 frontier 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 kontrolirana evaluacija modela" (zadane postavke, bez podešavanja i bez normalizacije troška).

Što su pronašli

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

Što ovo vjerojatno znači

Smanjenje halucinacija uglavnom nije pitanje izmišljanja još testova specifičnih za halucinacije. Riječ je o promjeni načina na koji glavni benchmarkovi boduju nesigurnost, tako da priznanje „ne znam” više nije kažnjeno. Dok se ljestvica ne promijeni, smanjivanje halucinacija i dalje će modelima oduzimati bodove točnosti i time biti obeshrabrivano — zato autori problem nazivaju „sociotehničkim”: dijelom je potrebna bolja metrika, a dijelom prihvaćanje te metrike na utjecajnim ljestvicama.

Što ovo *ne* dokazuje

- **Ne** pokazuje da otvorene rubrike rješavaju halucinacije u stvarnom svijetu. Potporni eksperiment mala je, namjerno **nekontrolirana** studija slučaja — četiri modela na zadanim postavkama, jedna tehnika ublažavanja, jedan test činjeničnih pitanja — zamišljena da pokaže *okret poticaja*, a ne rangira modele ili dokaže opću učinkovitost.
- **Ne** tvrdi da je bodovanje *jedini* uzrok. Pogreške u podacima za treniranje, stvarno teški problemi i nepoznati promptovi ostaju zasebni izvori.
- **Ne** podupire popularnu tvrdnju da su halucinacije neizbježne. Autori tvrde suprotno: sustav koji bi odgovarao samo na provjerljiva pitanja i inače govorio „ne znam” nikada ne bi halucinirao.
- **Ne** uklanja donju granicu predtreniranja — objašnjava je i omeđuje, a granica se odnosi na samouvjerene činjenične pogreške, ne na sve ponašanje modela.
- **Ne** pokazuje da su otvorene rubrike same po sebi *dovoljne*. Mijenjaju ono što evaluacija nagrađuje; nisu zamjena za dohvat podataka, alate ili bolje kalibrirane modele.

Koliko su dokazi jaki?

- Jezgra je **matematika** — formalne donje granice, ne mjerenja. Kao teorijski argument stoji na vlastitim pretpostavkama.
- Oslanja se na **namjerno pojednostavljene modele** problema; autori sami upozoravaju na „lažnu trihotomiju” u kojoj je svaki odgovor točan, pogrešan ili „ne znam”, i na idealizirano okruženje „proizvoljnih činjenica” korišteno za najčišću granicu.
- Pregled benchmarkova je **mali, kurirani uzorak** — deset utjecajnih evaluacija, ne iscrpan audit.
- Studija slučaja je **stvarna, ali ograničena**: četiri frontier modela, jedna tehnika ublažavanja, samo SimpleQA, zadane postavke, izričito „nije kontrolirana evaluacija”. To je dokaz koncepta za argument o poticajima, ne benchmark rezultat.
- Vrijedi navesti perspektivu autora: trojica od četvorice autora jesu ili su bili zaposlenici **OpenAI-ja**, a rad tvrdi da bi polje trebalo promijeniti način evaluacije modela. To je dobro argumentirana pozicija zainteresirane strane, ne neutralna vanjska revizija — treba je odvagnuti, ne odbaciti. (U prilog radu, kritiku jednako usmjerava prema vlastitim 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 djelomično samonametnutom — statistička donja granica koju možemo razumjeti, nadograđena poticajem koji smo sami odabrali. Šira pouka tiša je i korisnija: daljnji napredak u pouzdanosti mogao bi ovisiti jednako o tome *što mjerimo* kao i o tome što gradimo.

Sažetak bez ukrasa

Samouvjereni lažni odgovori jezičnih modela dolaze iz dva izvora. Prvi je statistički: kada činjenica nema uzorak koji se može naučiti, model treniran na oponašanju jezika ponekad će pogriješiti, a ta se donja granica može procijeniti, primjerice iz broja činjenica koje se u treningu pojavljuju samo jednom. Drugi su poticaji: gotovo svaki benchmark prema kojem se modeli rangiraju boduje „ne znam” isto kao pogrešan odgovor, pa pogađanje uvijek pobjeđuje — toliko da model koji griješi tri četvrtine vremena može nadmašiti pošteniji model koji odustaje. Prijedlog autora nije još jedan test halucinacija nego „otvorene rubrike”: navesti bodovanje u samom pitanju. U studiji slučaja na četiri frontier modela to okreće poticaj tako da se metoda za smanjenje halucinacija nagrađuje umjesto kažnjava. Riječ je o teorijskom radu s pregledom i malim, izričito nekontroliranim eksperimentom, recenziranom u časopisu *Nature*; rješenje je obećavajuće, ali još nije pokazano u velikom opsegu, a halucinacije prema argumentu nisu ni misteriozne ni strogo neizbježne.

Provjera bez uljepšavanja

Što rad pokazuje: Matematičku donju granicu prema kojoj je dio halucinacija prisiljen tijekom predtreniranja, najmanje „stopu jednokratnih pojavljivanja” za činjenice bez uzorka; pregled koji nalazi da većina vodećih benchmarkova ne daje nikakav bod za „ne znam”; i studiju slučaja s četiri modela u kojoj navođenje bodovanja u promptu, „otvorene rubrike”, omogućuje metodi za smanjenje halucinacija da pobijedi ondje gdje ju je obična točnost kažnjavala.

Što je vjerojatno, ali nije dokazano: Da bi otvorene rubrike, ako se ugrade u glavne benchmarkove, značajno smanjile halucinacije u modelima u uporabi. Potporni eksperiment malen je i izričito nekontroliran.

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

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

Koliko povjerenja treba imati opći čitatelj? Visoko da halucinacije nisu ni misteriozne ni strogo neizbježne i da glavni benchmarkovi trenutačno nagrađuju pogađanje. Umjereno da predloženo rješenje pomaže — sada ima stvaran dokaz koncepta, ali još ne i demonstraciju široke učinkovitosti 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