The Clean Paper

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

Prečo jazykové modely halucinujú – a prečo tento problém udržuje spôsob hodnotenia

Sebavedome prednášané nepravdy, ktorým hovoríme „halucinácie“, nie sú záhadnou závadou: niektoré sú štatistickým vedľajším produktom trénovania a problém pretrváva, pretože bežné benchmarky odmeňujú sebavedomé hádanie viac než poctivé „neviem“. Prípadová štúdia štyroch popredných modelov ukazuje, že uvedenie pravidiel bodovania priamo v zadaní – pomocou „explicitných kritérií“ – tento stimul obracia.

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

Prečo modely hádajú – a prečo sme ich to naučili

Spýtajte sa veľkého jazykového modelu na dátum narodenia málo známej osoby a môže odpovedať „7. marca“ s istotou niekoho, kto údaj práve číta z dokumentu – a pri troch pokusoch uviesť tri rôzne chybné dáta. Autori ukazujú práve také príklady: špičkové modely dostanú jednoduchú faktickú otázku – narodeniny človeka alebo rozvinutie málo známej skratky – a každý sebaisto ponúkne inú odpoveď, pričom všetky sú zle. Odvetvie tomu hovorí *halucinácia*, ako by išlo o poruchu vnímania. Prvým krokom je zbaviť tento jav tajomnosti.

Začnime tým, ako model vzniká. V prvej a najväčšej fáze tréningu sa z obrovského množstva textu učí, ako vyzerá plynulý jazyk. Zoberme si teraz údaj bez všeobecného vzoru – dátum narodenia konkrétneho človeka. Pokiaľ sa v trénovacích dátach objaví raz alebo vôbec, model nemá žiadny vzor, ktorý by zachytil: z jeho pohľadu je odpoveď ľubovoľná. Autori tento argument spresňujú pomocou staršej myšlienky Alana Turinga z iného odboru. Ak sa každý piaty dátum narodenia v dátach objaví len raz, možno očakávať, že model najmenej v jednom prípade z piatich chybuje – nie preto, že je pokazený, ale pretože v dátach nebolo nič, z čoho by sa mohol správnu odpoveď naučiť. Z rovnakého dôvodu sa modely takmer nikdy nemýlia pri hlavných mestách: také informácie sa opakujú neustále. Autori starostlivo dokazujú, že rozlíšiť pravdivú vetu od vierohodne znejúceho klamstva je samo o sebe náročná úloha a generovať *iba* pravdivé vety je prinajmenšom rovnako ťažké. Existuje teda neodstrániteľná dolná hranica chybovosti.

Základná myšlienka: ako spočítať to, čo sme ešte nevideli

Stojí za tým skutočne dômyselná myšlienka, ktorá je staršia než jazykové modely a zaslúži si podrobnejšie vysvetlenie.

Začnime sáčkom farebných guličiek. Neviete, koľko farieb obsahuje. Náhodne vytiahnete 100 guličiek a zaznamenáte výsledky: 40 červených, 25 modrých, 15 zelených, 5 žltých, 3 fialové, 2 oranžové - a potom **desať rôznych farieb, z ktorých každá sa objaví práve raz.**

Turing sa pri práci na úplne inom probléme spýtal: *aká je pravdepodobnosť, že budúca guličkа bude mať farbu, ktorú sme doposiaľ nevideli?* Farby, ktoré nikdy neboli vytiahnuté, spočítať nemožno - možno však spočítať tie, ktoré sa objavili presne raz, takzvané singletony. Metóda zvaná **Goodov-Turingov odhad** používa podiel týchto ojedinelých výskytov na odhad pravdepodobnosti skrytej v doposiaľ nepozorovaných farbách. Desať zo sto vytiahnutých guličiek malo farbu, ktorá sa objavila len raz, takže pravdepodobnosť, že ďalšia guličkа bude mať úplne novú farbu, je približne **10 / 100 = 10 %.**

Raz videné farby nie sú chyby. Sú meradlom našej nevedomosti: veľký počet jednotlivých výskytov je spôsob, ako sa nám snažia oznámiť, že svet obsahuje viac možností, z ktorých sme ešte nevzali vzorku.

Vymeňme teraz farby za dátumy narodenia a vrecko za tréningový text modelu. Predpokladajme, že medzi dátumami narodenia, ktoré model videl, sa **jedno z piatich vyskytlo práve raz.** Rovnaká metóda hovorí, že približne pätina pravdepodobnosti pripadá na dáta, s ktorými sa model nikdy nestretol – a dátum narodenia nemožno ničím nahradiť ani logicky *odvodiť*. Pri údajoch videných raz alebo vôbec tak model nemá dosť podkladov pre spoľahlivú voľbu a približne **jednu z piatich** takých odpovedí uvedie chybne. Žiadna dômyselnosť to nevyrieši: v

> dátach skrátka chýbal materiál, z ktorého by sa mohol učiť.
> To je celý argument v malom: podiel singletonov meria, akú veľkú časť sveta sa *z daných dát* nemožno naučiť, a stanovuje tak **dolnú hranicu** chybovosti. Vysvetľuje tiež, prečo sa model takmer nikdy nepomýli v hlavnom meste Francúzska – *Paríž* sa vyskytuje neustále, podiel ojedinelých výskytov sa blíži nule a materiálu na učenie je dostatok.

To vysvetľuje, odkiaľ halucinácie pochádzajú, nie však prečo pretrvávajú – prečo modely aj po ďalších fázach tréningu zameraných na užitočnosť a poctivosť ďalej blafujú, namiesto aby priznali neistotu. Prirovnanie autorov je bolestne presné. Predstavme si študenta, ktorý nepozná odpoveď pri skúške. Ak prázdna odpoveď znamená nula bodov a tip môže priniesť jeden, najlepšou stratégiou pre maximalizáciu skóre je hádať – sebavedome, priamo a bez dodatku „nie som si istý“. Študenti sa tomu prispôsobia. A rovnako tak modely, pretože ich hodnotíme podobne. Autori analyzovali benchmarky, v ktorých odvetvie skutočne súťaží, aj rebríčky, podľa ktorých sa modely dolaďujú. Takmer všetky dávajú za „neviem“ rovnaký počet bodov ako za chybnú odpoveď: nulu. Pri takom pravidle model, ktorý vždy háda, porazí inak totožný model, ktorý poctivo priznáva neistotu. Spôsobom bodovania ich k halucináciám doslova tlačíme.

Benchmarky môžu urobiť hádanie racionálnym. V typickom bodovacom systéme (vľavo) dostane nesprávna odpoveď aj úprimné „neviem“ nula bodov, takže hádanie môže jedine pomôcť. Keď sú v otázke uvedené pravidlá a zlá odpoveď je penalizovaná (vpravo), lepšou voľbou môže byť odmietnutie odpovede, keď si nie ste istí. To mení motiváciu vytvorenú testom, ale samo o sebe nerieši problém halucinácií.

Tento fragment stojí za zapamätanie, pretože odporuje typickým titulkom. Halucinácie sú často prezentované ako nevyhnutné, takmer mystické obmedzenia technológie. Práca spochybňuje obe

časti tohto rámca. Dolná hranica chybovosti pred tréningom nie je tajomstvom, ale jednoduchým štatistickým javom, ktorý je strojovému učeniu známy už desiatky rokov. Ich pretrvávanie tiež nie je nevyhnutné: je čiastočne spôsobené stimulmi, ktoré sme vytvorili a ktoré môžeme zmeniť. Systém, ktorý v prípade neistoty jednoducho odmietol reagovať, by vôbec nehalucinoval. Nasadené modely sa takto nesprávajú, pretože naše hodnotenie trestá odmietnutie.

Čo autori urobili

Práca sa skladá z troch častí. Prvá prináša matematický argument, podľa ktorého je počas predtréningu určitá miera halucinácií štatisticky vynútená: úloha „generovať iba správny text“ je prinajmenšom rovnako ťažká ako binárna klasifikácia „je táto veta pravdivá?“. Druhá časť – podporená prehľadom desiatich vplyvných benchmarkov – tvrdí, že bežné metriky presnosti odmeňujú hádanie viac ako zdržanlivosť. Tretia navrhuje nápravu a ukazuje ju v prípadovej štúdii: **explicitné hodnotiace kritériá**, pri ktorých sú pravidlá bodovania súčasťou samotnej otázky, napríklad: „správna odpoveď získa 1 bod, nesprávna −1; ak je vaša istota nižšia ako 50 %, nezodpovedajte.“ Model tak vie, kedy je poctivosť odmenená. Autori postup testujú na štyroch popredných modeloch – Gemini 3 Pro od Googlu, GPT-5 od OpenAI, Grok 4 od xAI a Claude Opus 4.5 od Anthropic – pomocou 4 326 faktických otázok zo sady SimpleQA. Výslovne uvádzajú, že ide o ilustratívnu prípadovú štúdiu, „nie o kontrolované porovnávacie hodnotenie modelov“: použili východiskové nastavenia, bez ladenia a bez normalizácie nákladov.

Čo našli

- **Predtréning vynucuje určitú chybovosť.** Miera sebaisto formulovaných neprávd má dolnú hranicu približne rovnú dvojnásobku chyby najlepšieho odvodeného klasifikátora „je toto tvrdenie pravdivé?“. Pri faktoch bez naučiteľného vzoru je touto medzou prinajmenšom *podiel singletonov* – teda faktov, ktoré sa v tréningu objavia práve raz. Určitá miera halucinácií je preto nevyhnutná aj pri dokonale čistých dátach.
- **Bodovanie priamo odmeňuje hádanie.** Pri obvyklom delení odpovedí na správne a nesprávne je optimálnou stratégiou nikdy sa nezdržať odpovede. Prehľad autorov ukazuje, že prevažná väčšina populárnych benchmarkov zaobchádza s „neviem“ jednoducho ako s chybou. Výrečný príklad sa týka vlastných modelov OpenAI: v SimpleQA hrubá presnosť mierne *zvýhodňuje* o4-mini, ktorý zodpovedá takmer vždy a vo viac ako troch štvrtinách prípadov sa mýli, oproti GPT-5-mini, ktorý robí oveľa menej chýb, pretože pri neistote častejšie neodpovie. Menej zdržanlivý model potom v rebríčku vyzerá lepšie.
- **Explicitné kritériá v prípadovej štúdii obracajú motiváciu.** Autori testujú jednoduchú metódu znižovania halucinácií: model odpovie dvakrát a odpovede sa vzdá, ak sa oba výsledky líšia. Pri štandardnom meraní presnosti metóda znižuje počet chýb, ale aj výsledné skóre, takže jej použitie hodnotenie znevýhodňuje. S explicitnými kritériami si tá istá metóda vedie lepšie pri všetkých štyroch modeloch v širokom rozmedzí sankcií. GPT-5-mini, ktorý prísna presnosť predtým trestala za zdržanlivosť, prekonáva o4-mini, keď sú pravidlá hodnotenia explicitné (n = 4 326 otázok na

model).

Čo to pravdepodobne znamená

Zníženie halucinácií nie je primárne o vymýšľaní viacerých testov špeciálne zameraných na halucinácie. Spôsob, akým mainstreamové benchmarky posudzujú neistotu, sa musí zmeniť, aby výrok „neviem“ nebol penalizovaný. Kým sa poradie nezmení, obmedzenie halucinácií bude stále stáť body presnosti, a tak zostaneme odradzovaní. Autori preto problém nazývajú „sociotechnickým“: sú potrebné lepšie metriky a zároveň treba presvedčiť vplyvné rebríčky, aby ich prijali.

Čo to *nedokazuje*

- Práca **neukazuje**, že explicitné kritériá opravujú halucinácie v aplikáciách v reálnom svete. Experiment podporujúci toto tvrdenie je malá, zámerne **nekontrolovaná** prípadová štúdia: štyri modely s východiskovým nastavením, jedna vybraná metóda a jeden test vecných otázok. Účelom je ukázať stimulačné obrátenie, nie klasifikovať modely alebo preukázať celkovú účinnosť.
- **Netvrdí**, že bodovanie je *jedinou* príčinou. Chyby v trénovacích dátach, skutočne ťažké problémy a neobvyklé príkazy zostávajú odlišnými zdrojmi.
- **Nepodporuje** populárne tvrdenie, že halucinácie sú nevyhnutné. Autori tvrdia opak: systém, ktorý zodpovedá iba vtedy, keď je možné odpoveď overiť, a inak hovorí „neviem“, by halucinovať nemusel.
- **Neodstraňuje** dolnú hranicu chýb pred tréningom – vyjasňuje ju a číselne obmedzuje, pričom sa táto hranica vzťahuje na uvádzané faktické chyby, nie na všetko správanie modelu.
- **Neukazuje**, že explicitné kritériá sú samé o sebe *dostatočné*. Menia to, čo hodnotenie odmeňuje, ale nenahrádzajú vyhľadávanie informácií, používanie nástrojov ani lepšie kalibrované modely.

Aké silné sú dôkazy?

- Jadrom práce je **matematika** – formálne dolné medze, nie meranie. Ako teoretický argument je platný na základe vlastných predpokladov.
- Je založený na **zámerne zjednodušených modeloch** problému. Sami autori poukazujú na „falošnú trichotómiu“ nakladania s každou odpoveďou ako správnu, nesprávnu alebo „neviem“ a na idealizovaný súbor „ľubovoľných faktov“ slúžiaci na získanie najčistejšej hranice.
- Prehľad benchmarkov tvorí **malá, vybraná vzorka** desiatich vplyvných hodnotení, nie vyčerpávajúci audit.
- Prípadová štúdia je **skutočná, ale obmedzená**: štyri hlavné modely, jedna metóda, iba SimpleQA, východiskové nastavenie a jasné vyhlásenie, že „toto nie je riadené hodnotenie“. Je to dôkaz princípu pre argument o stimuloch, nie výsledok hodnotenia.
- Stojí za to pomenovať pohľad autorov: traja zo štyroch pracujú alebo pracovali v **OpenAI** a článok

presviedča priemysel, aby zmenil spôsob hodnotenia modelov. Toto je dobre odôvodnený postoj zainteresovaných strán, nie neutrálne externé preskúmanie – mal by byť zvážený, nie zamietnutý. Výhodou autorov je, že kritizujú svoje vlastné modely, o4-mini a GPT-5-mini, rovnako ochotne ako kritizujú modely iných spoločností.

Prečo na tom záleží

Práca mení spôsob uvažovania o veľmi medializovanom probléme. „Halucinácia“ je obvykle prezentovaná ako strašidelná chyba alebo nepriechodný múr; tu sa stáva niečím obyčajným a čiastočne samovoľne vyvolaným – pochopiteľnou štatistickou hranicou, na ktorú je uvalený nami zvolený systém stimulov. Širšie poučenie je pokojnejšie a užitočnejšie: pokračujúci pokrok v spoľahlivosti môže závisieť od toho, čo meriame, ako aj od toho, čo budujeme.

Stručné zhrnutie

Falošné odpovede jazykových modelov pravdepodobne pochádzajú z dvoch zdrojov. Prvý je štatistický: ak za faktom nie je žiadny vzor, model vycvičený na napodobňovanie jazyka ho niekedy uhádne nesprávne a dolnú hranicu chybovosti možno odhadnúť napríklad podľa počtu faktov, ktoré sa v tréningu objavia len raz. Druhým zdrojom sú stimuly: takmer každý benchmark používaný na zostavenie rebríčka udeľuje za „neviem“ rovnako bodov ako za zlú odpoveď, takže sa hádanie vždy oplatí – aj model, ktorý sa v troch štvrtinách prípadov mýli, môže poraziť poctivejší model, ktorý sa odpovede zdrží. Návrh autorov nie je ďalším testom halucinácií, ale použitím „explicitných kritérií“: pravidlá bodovania sa vložia priamo do zadania. V prípadovej štúdii štyroch hlavných modelov sa tým motivácia obrátila, takže metóda obmedzujúca halucinácie bola odmeňovaná, nie trestaná. Ide o recenzovaný článok v Nature, ktorý kombinuje teóriu, prehľad a malý, výslovne nekontrolovaný experiment. Riešenie je sľubné, ale jeho účinnosť zatiaľ nebola vo veľkom meradle preukázaná; autori tvrdia, že halucinácie nie sú ani záhadné, ani striktne nevyhnutné.

Kontrola bez príkras

Čo práca ukazuje: Matematickú dolnú hranicu, podľa ktorej je časť halucinácií počas predtréningu nevyhnutná – pri faktoch bez vzoru zodpovedá najmenej podielu singletonov; prehľad ukazujúci, že väčšina popredných benchmarkov neprideľuje body za „neviem“; a štvormodelová štúdia, v ktorej uvedenie bodovania priamo v zadaní pomocou explicitných kritérií zvýhodnilo metódu obmedzujúcu halucinácie, aj keď ju penalizovala jednoduchá presnosť.

Čo je vierohodné, ale nepreukázané: Že zavedenie explicitných kritérií do mainstreamových benchmarkov by výrazne znížilo halucinácie v implementovaných modeloch. Podporný experiment je malý a zjavne nekontrolovaný.

Čo práca neukazuje: Že halucinácie sú nevyhnutné – tvrdí opak; že bodovanie je jedinou príčinou; že halucinácie možno úplne odstrániť; ani to, že prípadová štúdia predstavuje rebríček štyroch modelov.

Hlavné obmedzenia: Zámerne zjednodušené delenie na správne, nesprávne a „neviem“ odpovede, ktoré autori nazývajú „falošná trichotómia“; malý, vybraný prehľad desiatich benchmarkov; nekontrolovaná štúdia štyroch modelov, jednej metódy a jedného testu s východiskovým nastavením; a skutočnosť, že argument pre hodnotenie v celom odvetví pochádzal predovšetkým od výskumníkov OpenAI.

Koľko dôvery by mal mať nešpecializovaný čitateľ? Vysokú v záver, že halucinácie nie sú ani záhadné, ani striktne nevyhnutné a že bežné benchmarky dnes odmeňujú hádanie. Strednú v tom, že navrhovaná náprava pomôže: existuje skutočný dôkaz princípu, ale účinnosť vo veľkom meradle zatiaľ preukázaná nebola.

Zdroje

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