The Clean Paper

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

Proč jazykové modely halucinují – a proč tento problém udržuje způsob hodnocení

Sebevědomě pronášené nepravdy, kterým říkáme „halucinace“, nejsou záhadnou závadou: některé jsou statistickým vedlejším produktem trénování a problém přetrvává, protože běžné benchmarky odměňují sebevědomé hádání více než poctivé „nevím“. Případová studie čtyř předních modelů ukazuje, že uvedení pravidel bodování přímo v zadání – pomocí „explicitních kritérií“ – tuto pobídku obrací.

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

Proč modely hádají – a proč jsme je to naučili

Zeptejte se velkého jazykového modelu na datum narození málo známé osoby a může odpovědět „7. března“ s jistotou někoho, kdo údaj právě čte z dokumentu – a při třech pokusech uvést tři různá chybná data. Autoři ukazují právě takové příklady: špičkové modely dostanou jednoduchou faktickou otázku – narozeniny člověka nebo rozvinutí málo známé zkratky – a každý sebejistě nabídne jinou odpověď, přičemž všechny jsou špatně. Odvětví tomu říká *halucinace*, jako by šlo o poruchu vnímání. Prvním krokem je zbavit tento jev tajemnosti.

Začněme tím, jak model vzniká. V první a největší fázi tréninku se z obrovského množství textu učí, jak vypadá plynulý jazyk. Vezměme si nyní údaj bez obecného vzoru – datum narození konkrétního člověka. Pokud se v trénovacích datech objeví jednou nebo vůbec, model nemá žádný vzor, který by zachytil: z jeho pohledu je odpověď libovolná. Autoři tento argument zpřesňují pomocí starší myšlenky Alana Turinga z jiného oboru. Jestliže se každé páté datum narození v datech objeví jen jednou, lze očekávat, že model nejméně v jednom případě z pěti chybuje – ne proto, že je porouchaný, ale protože v datech nebylo nic, z čeho by se mohl správnou odpověď naučit. Ze stejného důvodu modely téměř nikdy nechybují u hlavních měst: takové informace se opakují neustále. Autoři pečlivě dokazují, že rozlišit pravdivou větu od věrohodně znějící lži je samo o sobě obtížný úkol a generovat *pouze* pravdivé věty je přinejmenším stejně těžké. Existuje tedy neodstranitelná dolní mez chybovosti.

Základní myšlenka: jak spočítat to, co jsme ještě neviděli

To je založeno na skutečně důmyslné myšlence, která je starší než jazykové modely a stojí za podrobné prozkoumání.

Začněme sáčkem barevných kuliček. Nevíte, kolik barev obsahuje. Náhodně vytáhnete 100 kuliček a zaznamenáte výsledky: 40 červených, 25 modrých, 15 zelených, 5 žlutých, 3 fialové, 2 oranžové – a pak **deset různých barev, z nichž každá se objeví právě jednou.**

Turing se při práci na zcela jiném problému zeptal: *jaká je pravděpodobnost, že příští kulička bude mít barvu, kterou jsme dosud neviděli?* Barvy, které nikdy nebyly vytaženy, spočítat nelze – lze však spočítat ty, které se objevily přesně jednou, takzvané singletony. Metoda zvaná **Goodův–Turingův odhad** používá podíl těchto ojedinělých výskytů k odhadu pravděpodobnosti skryté v dosud nepozorovaných barvách. Deset ze sta vytažených kuliček mělo barvu, která se objevila jen jednou, takže pravděpodobnost, že další kulička bude mít zcela novou barvu, činí přibližně **10 / 100 = 10 %.**

Jednou viděné barvy nejsou *chyby*. Jsou měřítkem naší nevědomosti: velký počet jednotlivých výskytů je způsob, jak se nám snaží sdělit, že svět obsahuje více možností, které jsme ještě nenačerpali.

Vyměňme nyní barvy za data narození a sáček za tréninkový text modelu. Předpokládejme, že mezi daty narození, která model viděl, se **jedno z pěti vyskytlo právě jednou.** Stejná metoda říká, že přibližně pětina pravděpodobnosti připadá na data, s nimiž se model nikdy nesetkal – a datum narození nelze ničím nahradit ani logicky *odvodit*. U údaje spatřeného jednou nebo vůbec tak model nemá dost podkladů pro spolehlivou volbu a přibližně **jednu z pěti** takových odpovědí uvede chybně. Žádná důmyslnost to nevyřeší: v datech zkrátka chyběl materiál, z něhož by se mohl učit.

> To je celý argument v malém: podíl singletonů měří, jak velkou část světa se *z daných dat* nelze naučit, a stanovuje tak **dolní mez** chybovosti. Vysvětluje také, proč se model téměř nikdy nesplete v hlavním městě Francie – *Paříž* se vyskytuje neustále, podíl ojedinělých výskytů se blíží nule a materiálu k učení je dostatek.

To vysvětluje, odkud halucinace pocházejí, nikoli však proč *přetrvávají* – proč modely i po dalších fázích tréninku zaměřených na užitečnost a poctivost dál blafují, místo aby přiznaly nejistotu. Přirovnání autorů je bolestně přesné. Představme si studenta, který nezná odpověď u zkoušky. Jestliže prázdná odpověď znamená nula bodů a tip může přinést jeden, nejlepší strategií pro maximalizaci skóre je hádat – sebevědomě, přímo a bez dodatku „nejsem si jistý“. Studenti se tomu přizpůsobí. A stejně tak modely, protože je hodnotíme podobně. Autoři analyzovali benchmarky, v nichž odvětví skutečně soutěží, i žebříčky, podle kterých modely dolaďuje. Téměř všechny dávají za „nevím“ stejně bodů jako za chybnou odpověď: nulu. Při takovém pravidle model, který vždy hádá, porazí jinak totožný model, jenž poctivě přiznává nejistotu. Způsobem bodování je k halucinacím doslova tlačíme.

Benchmarky mohou učinit hádání racionálním. V typickém bodovacím systému (vlevo) má jak nesprávná odpověď, tak upřímné „nevím“ nulu, takže hádání může jedině pomoci. Když jsou v otázce uvedena pravidla a špatná odpověď je penalizována (správně), lepší volbou může být odmítnutí odpovědi, když si nejste jisti. To mění motivaci vytvořenou testem, ale samo o sobě neřeší problém halucinací.

Tento fragment stojí za zapamatování, protože odporuje typickým titulkům. Halucinace jsou často prezentovány jako nevyhnutelné, téměř mystické omezení technologie. Práce zpochybňuje obě definice. Spodní mez pro chyby před tréninkem není tajemstvím, ale je to jednoduchá statistická

chyba, která je strojovému učení známá již desítky let. Jejich přetrvávání také není nevyhnutelné: je částečně způsobeno pobídkami, které jsme vytvořili a můžeme je změnit. Systém, který v případě nejistoty jednoduše odmítl reagovat, by vůbec nehalucinoval. Nasazené modely se takto nechovají, protože naše hodnocení trestá odmítnutí.

Co autoři udělali

Práce se skládá ze tří částí. První přináší matematický argument, podle něhož je během předtréninku určitá míra halucinací statisticky vynucená: úloha „generovat pouze správný text“ je přinejmenším stejně obtížná jako binární klasifikace „je tato věta pravdivá?“. Druhá část – podpořená přehledem deseti vlivných benchmarků – tvrdí, že běžná měřítka přesnosti odměňují hádání více než zdrženlivost. Třetí navrhuje nápravu a ukazuje ji v případové studii: **explicitní hodnoticí kritéria**, při nichž jsou pravidla bodování součástí samotné otázky, například: „správná odpověď získá 1 bod, nesprávná −1; pokud je vaše jistota nižší než 50 %, neodpovídejte.“ Model tak ví, kdy je poctivost odměněna. Autoři postup testují na čtyřech předních modelech – Gemini 3 Pro od Googlu, GPT-5 od OpenAI, Grok 4 od xAI a Claude Opus 4.5 od Anthropic – pomocí 4 326 faktických otázek ze sady SimpleQA. Výslovně uvádějí, že jde o ilustrativní případovou studii, „nikoli o kontrolované srovnávací hodnocení modelů“: použili výchozí nastavení, bez ladění a bez normalizace nákladů.

Co našli

- **Předtrénink vynucuje určitou míru chyb.** Míra sebejistě formulovaných nepravd má dolní mez přibližně rovnou dvojnásobku chyby nejlepšího odvozeného klasifikátoru „je toto tvrzení pravdivé?“. U faktů bez naučitelného vzoru je touto mezí přinejmenším *podíl singletonů* – tedy faktů, které se v tréninku objeví právě jednou. Určitá míra halucinací je proto nevyhnutelná i při dokonale čistých datech.
- **Bodování přímo odměňuje hádání.** Při obvyklém dělení odpovědí na správné a nesprávné je optimální strategií nikdy se nezdržet odpovědi. Přehled autorů ukazuje, že naprostá většina populárních benchmarků zachází s „nevím“ jednoduše jako s chybou. Výmluvný příklad se týká vlastních modelů OpenAI: v SimpleQA hrubá přesnost mírně *zvýhodňuje* o4-mini, který odpovídá téměř vždy a ve více než třech čtvrtinách případů se mýlí, oproti GPT-5-mini, jenž dělá mnohem méně chyb, protože při nejistotě častěji neodpoví. Méně zdrženlivý model pak v žebříčku vypadá lépe.
- **Explicitní kritéria v případové studii obracejí motivaci.** Autoři testují jednoduchou metodu snižování halucinací: model odpoví dvakrát a odpovědi se vzdá, pokud se oba výsledky liší. Při standardním měření přesnosti metoda snižuje počet chyb, ale také výsledné skóre, takže její použití hodnocení znevýhodňuje. S explicitními kritérii si tatáž metoda vede lépe u všech čtyř modelů v širokém rozmezí sankcí. GPT-5-mini, který přísná přesnost dříve trestala za zdrženlivost, překonává o4-mini, když jsou pravidla hodnocení explicitní (n = 4 326 otázek na model).

Co to pravděpodobně znamená

Snížení halucinací není primárně o vymýšlení více testů speciálně zaměřených na halucinace. Způsob, jakým mainstreamové benchmarky posuzují nejistotu, se musí změnit, aby výrok „nevím“ nebyl penalizován. Dokud se pořadí nezmění, omezení halucinací bude stále stát body přesnosti, a tak zůstaneme odrazováni. Autoři proto problém nazývají „socio-technickým“: je potřeba jak lepší opatření, tak vlivné žebříčky jsou k jeho přijetí přesvědčovány.

Co to *nedokazuje*

- Práce **neukazuje**, že explicitní kritéria opravují halucinace v aplikacích v reálném světě. Experiment podporující toto tvrzení je malá, záměrně **nekontrolovaná** případová studie: čtyři modely s výchozím nastavením, jedna vybraná metoda a jeden test věcných otázek. Účelem je ukázat *pobídkové obrácení*, nikoli klasifikovat modely nebo prokázat celkovou účinnost.
- **Netvrdí**, že bodování je *jedinou* příčinou. Chyby v trénovacích datech, skutečně obtížné problémy a neobvyklé příkazy zůstávají odlišnými zdroji.
- **Nepodporuje** populární tvrzení, že halucinace jsou nevyhnutelné. Autoři tvrdí opak: systém, který odpovídá pouze tehdy, když lze odpověď ověřit, a jinak říká „nevím“, by halucinovat nemusel.
- **Neodstraňuje** dolní mez chyb před tréninkem – vyjasňuje ji a číselně omezuje, přičemž se tato mez pravděpodobně týká vykazovaných faktických chyb, nikoli veškerého chování modelu.
- **Neukazuje**, že explicitní kritéria jsou sama o sobě *dostatečná*. Mění to, co hodnocení odměňuje, ale nenahrazují vyhledávání informací, používání nástrojů ani lépe kalibrované modely.

Jak silné jsou důkazy?

- Jádrem práce je **matematika** - formální dolní meze, nikoli měření. Jako teoretický argument je platný na základě vlastních předpokladů.
- Je založen na **záměrně zjednodušených modelech** problému. Sami autoři poukazují na „falešnou trichotomii“ nakládání s každou odpovědí jako správnou, nesprávnou nebo „nevím“ a na idealizovaný soubor „libovolných faktů“ sloužící k získání nejčistší hranice.
- Referenční přezkum je **malý, vybraný vzorek** deseti posouzení dopadu, nikoli vyčerpávající audit.
- Případová studie je **skutečná, ale omezená**: čtyři hlavní modely, jedna metoda, pouze SimpleQA, výchozí nastavení a jasné prohlášení, že „toto není řízené hodnocení“. Toto je důkaz koncepce pro argument pobídek, nikoli výsledek hodnocení.
- Stojí za to pojmenovat pohled autorů: tři ze čtyř pracují nebo pracovali v **OpenAI** a článek přesvědčuje průmysl, aby změnil způsob hodnocení modelů. Toto je dobře odůvodněný postoj zúčastněných stran, nikoli neutrální externí přezkum – měl by být zvážen, nikoli zamítnut. Výhodou autorů je, že kritizují své vlastní modely, o4-mini a GPT-5-mini, stejně ochotně jako kritizují modely jiných společností.

Proč na tom záleží

Práce mění způsob uvažování o velmi medializovaném problému. „Halucinace“ je obvykle prezentována jako strašidelná chyba nebo neprůchodná zeď; zde se stává něčím obyčejným a částečně samovolně vyvolaným – pochopitelnou statistickou hranicí, na kterou je uvalen námi zvolený systém pobídek. Širší poučení je klidnější a užitečnější: pokračující pokrok ve spolehlivosti může záviset jak na tom, co měříme, tak na tom, co budujeme.

Stručné shrnutí

Falešné odpovědi jazykových modelů pravděpodobně pocházejí ze dvou zdrojů. První je statistický: pokud za faktem není žádný vzor, model vycvičený k napodobování jazyka jej někdy zamění a dolní mez chybovosti lze odhadnout například podle počtu faktů, které se v tréninku objeví jen jednou. Druhým zdrojem jsou pobídky: téměř každý benchmark používaný k sestavení žebříčku uděluje za „nevím“ stejně bodů jako za špatnou odpověď, takže se hádání vždy vyplatí – i model, který se ve třech čtvrtinách případů mýlí, může porazit poctivější model, jenž se odpovědi zdrží. Návrh autorů není dalším testem halucinací, ale použitím „explicitních kritérií“: pravidla bodování se vloží přímo do zadání. V případové studii čtyř hlavních modelů se tím pobídka obrátila, takže metoda omezující halucinace byla odměňována, nikoli trestána. Jde o recenzovaný článek v *Nature*, který kombinuje teorii, přehled a malý, výslovně nekontrolovaný experiment. Řešení je slibné, ale jeho účinnost zatím nebyla ve velkém měřítku prokázána; autoři tvrdí, že halucinace nejsou ani záhadné, ani striktně nevyhnutelné.

Kontrola bez příkras

Co práce ukazuje: Matematická dolní mez, při které je část halucinací vynucena během předtréninku – u bezvzorových faktů se alespoň rovná procentu singletonů; revize ukazující, že většina předních benchmarků nepřiděluje body za „nevím“; a čtyřmodelová studie, ve které poskytnutí skóre v příkazu, to znamená pomocí explicitních kritérií, způsobilo vítězství metody omezující halucinace, i když ji penalizovala jednoduchá přesnost.

Co je věrohodné, ale neprokázané: Že zavedení explicitních kritérií do mainstreamových benchmarků by výrazně snížilo halucinace v implementovaných modelech. Podpůrný experiment je malý a zjevně nekontrolovaný.

Co práce neukazuje: Že halucinace jsou nevyhnutelné, tvrdí opak; že bodování je jedinou příčinou; že halucinace mohou být zcela odstraněny; že případová studie je žebříček čtyř modelů.

Hlavní omezení: Záměrně zjednodušené dělení na správné, nesprávné a „nevím“ odpovědi, které autoři nazývají „falešná trichotomie“; malý, vybraný přehled deseti benchmarků; nekontrolované studium čtyř modelů, jedné metody a jednoho testu s výchozím nastavením; a skutečnost, že argu-

ment pro hodnocení v celém odvětví pocházel především od výzkumníků OpenAI.

Kolik důvěry by měl mít nespecializovaný čtenář? Vysokou v závěr, že halucinace nejsou ani záhadné, ani striktně nevyhnutelné a že běžné benchmarky dnes odměňují hádání. Střední v to, že navrhovaná náprava pomůže: existuje skutečný důkaz principu, ale účinnost ve velkém měřítku zatím prokázána nebyla.

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