The Clean Paper

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

Зошто јазичните модели халуцинираат — и зошто начинот на кој ги оценуваме го одржува проблемот

Самоуверените неточности што ги нарекуваме „халуцинации“ не се мистериозен дефект: дел се статистички нуспроизвод на обуката, а опстојуваат затоа што главните benchmark-и наградуваат самоуверено погодување повеќе од искрено „не знам“. Case study со четири frontier модели покажува дека наведувањето на правилата за бодирање во prompt-от („open rubrics“) го превртува тој поттик.

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

Зошто моделите погодуваат — и зошто ние ги научивме да го прават тоа

Прашајте голем јазичен модел за роденденот на непозната личност и може да одговори „7 март“ со мирна самоувереност како да ја чита датата од картичка — и да биде погрешен, трипати по ред, секојпат со различна дата. Авторите даваат токму вакви примери: водечки модели добиваат едноставно фактичко прашање — нечиј роденден или значењето на редок акроним — и секој самоуверено измислува друг одговор, сите погрешни. Индустријата ова го нарекува *халуцинација*, збор што звучи како дефект на перцепцијата. Првиот потег на трудот е да ја отстрани мистиката.

Да почнеме од тоа како се создава модел. Во првата и најголема фаза на обука, тој практично учи како изгледа природен и течен јазик читајќи огромна количина текст. Сега земете факт зад кој нема образец — точниот роденден на една конкретна личност. Ако таа дата се појавила еднаш во податоците за обука, или никогаш, нема образец за кој систем што учи обрасци може да се фати: од гледна точка на моделот, одговорот е произволен. Авторите ова го формализираат со стара идеја, потекната од Alan Turing за сосема друг проблем: ако една петтина од родендените во податоците се појавуваат само еднаш, треба да очекуваме моделот да згреши најмалку една петтина од нив — не затоа што е „расипан“, туку затоа што едноставно немало доволно податоци од кои може да научи. (Од истата причина моделите речиси никогаш не го промашуваат главниот град на држава: тие факти се појавуваат постојано.) Авторите внимателно тврдат дека разликувањето вистинита од убедливо лажна реченица и самото е тежок проблем, а создавањето *само* вистинити реченици е најмалку толку тешко. Одредена долна граница на грешка е вградена во задачата.

Идејата под ова: како да го процениме она што сè уште не сме го виделе

Во основата стои навистина паметна идеја — постара од јазичните модели и вредна сама по себе.

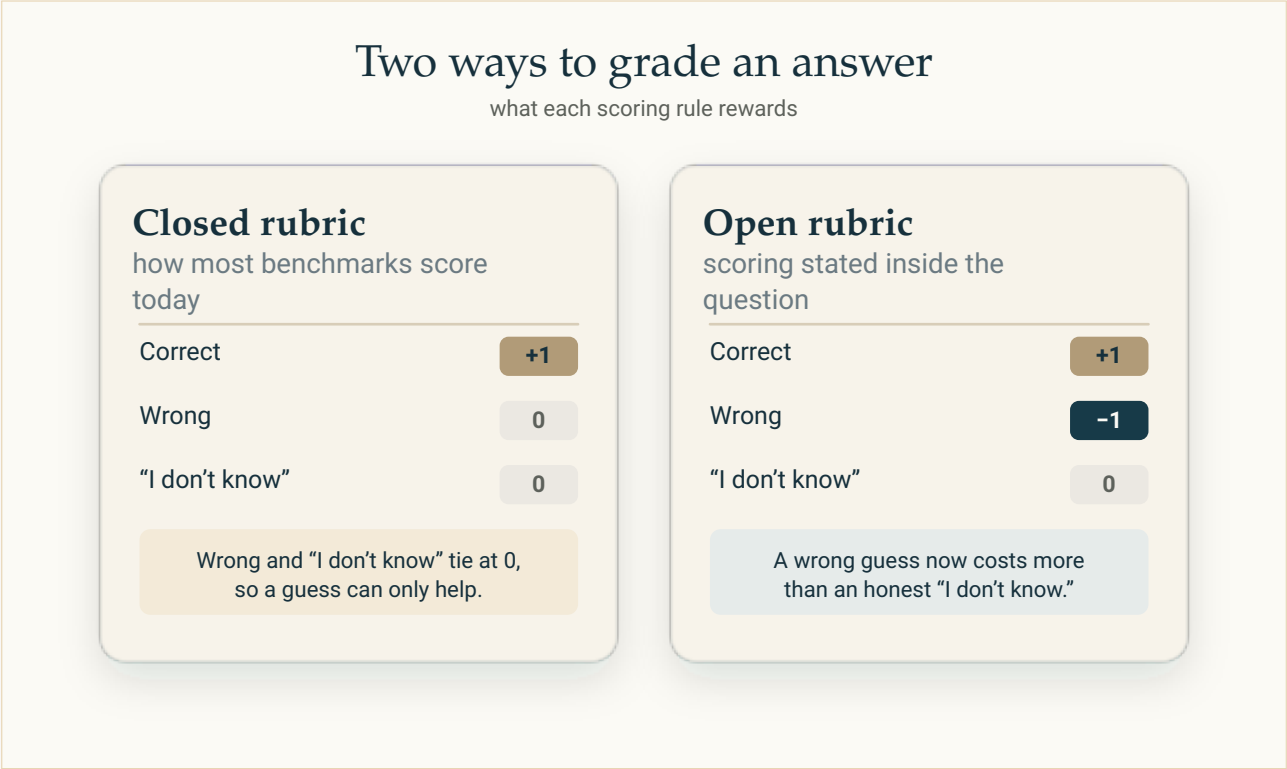
Замислете торба со обоени топчиња. Не знаете колку бои има. Извлекувате 100 топчиња, едно по едно, и броите: црвени 40, сини 25, зелени 15, жолти 5, виолетови 3, портокалови 2 — а потоа **десет различни бои што се појавуваат точно по еднаш**.

Прашањето со кое Turing се соочил во друг контекст било: *колкава е шансата следното топче да биде боја што досега воопшто не сте ја виделе?* Не можете директно да го изброите она што никогаш не сте го извлекле — но можете да ги изброите боите што сте ги виделе точно еднаш, „случаи што се појавуваат само еднаш“. Трикот, наречен **Good–Turing проценка**, вели дека уделот на извлекувањата што се singletons ја проценува веројатноста скриена во сè уште невидените бои. Десет од сто извлекувања биле бои што се појавиле еднаш, па шансата следното да биде сосема нова боја е околу **10 / 100 = 10%**.

Тие еднаш видени бои не се *грешки*. Тие се мерка на сопственото незнаење: многу бои што се појавуваат по еднаш му кажуваат на примерокот дека светот содржи

> уште нешта што едноставно не сте ги извлекле.
> **Сега заменете ги боите со родендени,** а торбата со текстот за обука. Ако, меѓу родендените што ги видел моделот, **еден од пет се појавува точно еднаш**, истиот трик кажува дека приближно една петтина од веројатноста лежи во родендени што моделот практично никогаш не ги научил — а за роденден нема образец што може да се пресмета. Дата видена еднаш или никогаш е нешто за што моделот нема основа да ја тежи веројатноста и ќе греши приближно кај **еден од пет** такви случаи. Никаква „паметност“ не го поправа тоа: немало доволно информација за учење.
> Тоа е целиот аргумент во минијатура: стапката на singletons мери колкав дел од светот не може да се научи *од овие податоци*, а тоа поставува **долна граница** под грешките. И објаснува зошто модел речиси никогаш не го промашува главниот град — *Paris* се појавува постојано, singleton-стапката е речиси нула и има многу што да се научи.

Тоа објаснува од каде доаѓаат халуцинациите. Не објаснува зошто тие *опстојуваат* — зошто моделите, по дополнителната обука што треба да ги направи корисни и искрени, сè уште блефираат наместо да признаат сомнеж. Тука аналогијата во трудот е непријатно точна. Замислете ученик на испит што не го знае одговорот. Ако празно поле носи нула, а погодување може да донесе еден поен, стратегијата што ја максимизира оценката е да погоди — конкретно и самоуверено, никогаш „не сум сигурен“. Учениците го учат тоа. Се покажува дека и моделите го учат — затоа што и нив ги оценуваме така. Авторите ги прегледале benchmark-ите на кои полето навистина се натпреварува, leaderboard-ите за кои моделите се оптимизираат, и откриле дека речиси сите му даваат на „не знам“ ист резултат како на погрешен одговор: нула. Под такво правило, модел што секогаш погодува ќе го надмине инаку идентичниот модел што искрено ја означува неизвесноста. Во прилично буквална смисла, нашиот систем на бодирање ги поттикнува да блефираат.



Benchmark-ите можат да направат погодувањето рационална стратегија. Со вообичаено бодирање (лево), погрешен одговор и искрено „не знам“ носат нула — па погодувањето може само да помогне. Ако правилата се наведат во самото прашање и погрешниот одговор се казнува (десно), воздржувањето кога моделот не е сигурен може да стане подобар избор. Тоа менува што наградува тестот; само по себе не го решава проблемот со халуцинации.

Original diagram — The Clean Paper CC BY 4.0

Овој дел вреди да се запамети бидејќи оди спротивно на вообичаениот наслов. Халуцинациите често се претставуваат како неизбежна, речиси мистична граница на технологијата. Трудот го оспорува и „мистичното“ и „неизбежното“. Долната граница од pretraining не е мистерија — тоа е обична статистичка грешка, тип што машинското учење го разбира со децении. А опстојувањето не е целосно неизбежно — делумно е поттик што самите сме го изградиле и можеме да го смениме. Систем што едноставно одбива да одговори кога не е сигурен не би халуцинирал; причината зошто применетите модели не се однесуваат така е што нашите табели со резултати го казнуваат воздржувањето.

Што направиле авторите

Трудот има три дела. Прво, математички аргумент дека одредена количина халуцинација е статистички присилена во pretraining, со покажување дека задачата „генерирај само валиден текст“ е најмалку толку тешка колку бинарната класификација „дали оваа изјава е валидна?“. Второ, аргумент — поддржан со преглед на десет влијателни benchmark-и — дека вообичаените метрики на точност го наградуваат погодувањето

наместо воздржувањето. Трето, предложена поправка и case study што ја тестира: **open-rubric** евалуации, каде бодирањето е наведено во самото прашање (на пример, „точен одговор носи 1, погрешен −1, па воздржи се ако си помалку од 50% сигурен“), така што моделот знае кога искреноста се наградува. Ова го тестираат на четири frontier модели — Google's Gemini 3 Pro, OpenAI's GPT-5, xAI's Grok 4 и Anthropic's Claude Opus 4.5 — со 4.326 фактички прашања од SimpleQA. Јасно велат дека case study е илустративен, „не контролирана евалуација меѓу модели“: default поставки, без tuning и без нормализација на трошокот.

Што откриле

- **Pretraining присилува одредена грешка.** Стапката со која модел создава самоуверени лажни тврдења има долна граница поврзана со грешката на најдобриот класификатор „дали оваа изјава е валидна?“ изграден од него. За факти без научлив образец, таа граница е најмалку *singleton-стапката* — уделот на факти што се појавуваат точно еднаш во обуката. Некои халуцинации се неизбежни дури и со совршено чисти податоци.
- **Бодирањето конкретно го наградува погодувањето.** Со обично „точно/погрешно“ бодирање, никогаш да не се воздржуваш е оптимална стратегија, а прегледот на авторите покажува дека огромното мнозинство популарни benchmark-и го оценуваат „не знам“ едноставно како погрешно. Впечатлив пример од нивните сопствени модели: на SimpleQA, суровата точност благо го фаворизира o4-mini на OpenAI — кој одговара речиси секогаш и греши повеќе од три четвртини од времето — пред GPT-5-mini, кој прави многу помалку грешки бидејќи се воздржува кога не е сигурен. Поризичниот модел изгледа подобро на scoreboard-от.
- **Open rubrics го превртуваат поттикот во нивниот case study.** Тестираат едноставна мерка против халуцинации: моделот да одговори двапати и да се воздржи ако двата одговора не се согласуваат. Под стандардна точност, мерката ги намалува грешките *но и точноста*, па метриката ја казнува. Под open rubrics, истата мерка е подобра за сите четири модели низ опсег на казни; а GPT-5-mini — казнет од суровата точност затоа што се воздржува — го надминува o4-mini кога правилата за бодирање се експлицитно наведени (n = 4.326 прашања по модел).

Што најверојатно значи ова

Намалувањето на халуцинациите веројатно не е главно прашање на создавање уште повеќе специјални тестови за халуцинации. Повеќе е прашање на тоа како главните benchmark-и ја оценуваат неизвесноста, така што признавањето „не знам“ да не се казнува. Додека scoreboard-от не се смени, намалувањето на халуцинациите ќе продолжи да чини поени на accuracy и ќе се обесхрабрува — затоа авторите проблемот го нарекуваат „социотехнички“: дел е подобра метрика, дел е убедување на

влијателните leaderboard-и да ја усвојат.

Што ова не докажува

- **Не** покажува дека open rubrics ги решаваат халуцинациите во реална употреба. Поддржувачкиот експеримент е мал и намерно **неконтролиран** case study — четири модели со default поставки, една избрана мерка и еден фактички QA-тест — направен да го покаже *превртувањето на поттикот*, не да рангира модели или да докаже општа ефикасност.
- **Не** тврди дека бодирањето е *единствената* причина. Грешки во податоците за обука, навистина тешки проблеми и непознати prompt-ови остануваат посебни извори.
- **Не** ја поддржува популарната теза дека халуцинациите се неизбежни. Авторите тврдат спротивното: систем што одговара само на проверливи прашања и во спротивно вели „не знам“ не би халуцинирал.
- **Не** ја отстранува долната граница од pretraining — ја објаснува и ограничува, а границата се однесува на самоуверени фактички грешки, не на целото однесување на моделот.
- **Не** покажува дека open rubrics се *доволни* сами по себе. Тие менуваат што наградува евалуацијата; не се замена за retrieval, алатки или подобро калибрирани модели.

Колку се убедливи доказите?

- Јадрото е **математика** — формални долни граници, не емпириски мерења. Како теоретски аргумент е цврсто во рамките на претпоставките.
- Се потпира на **намерно поедноставени модели** на проблемот; самите автори ја нагласуваат „лажната трихотомија“ во која секој одговор е или точен, или погрешен, или „не знам“, и идеализираниот свет на „произволни факти“ користен за најчистата граница.
- Прегледот на benchmark-и е **мал, куриран примерок** — десет влијателни евалуации, не исцрпен аудит.
- Case study е **реален, но ограничен**: четири frontier модели, една мерка, само SimpleQA, default поставки, експлицитно „не контролирана евалуација“. Тоа е proof of concept за аргументот за поттик, не benchmark-резултат.
- Вреди да се именува и гледната точка: тројца од четворицата автори се или биле вработени во **OpenAI**, а трудот тврди дека полето треба да го промени начинот на оценување модели. Тоа е добро аргументирана позиција од заинтересирана страна, не неутрален надворешен преглед — треба да се измери, не да се отфрли. (Во корист на трудот, критиката еднакво ги опфаќа и сопствените модели o4-mini и GPT-5-mini.)

Зошто е важно

Трудот прерамкува еден многу хиперболизиран проблем. „Халуцинација“ често се продава или како чуден дефект или како непоместлив sид; овој труд ја прави обична и делумно самонаметната — статистичка долна граница што можеме да ја разбереме, поставена врз поттик што сами сме го избрале. Пошироката поука е потивка и покорисна: идниот напредок во сигурноста можеби ќе зависи исто толку од *тоа што го мериме* колку од тоа што го градиме.

Кратко резиме

Самоуверените лажни одговори од јазични модели доаѓаат од две места. Првото е статистичко: кога факт нема образец што може да се научи, модел обучен да имитира јазик понекогаш ќе го погоди погрешно, а долната граница на таа грешка може да се процени — на пример од тоа колку факти се појавуваат само еднаш во обуката. Второто се поттиците: речиси секој benchmark на кој моделите се рангираат го оценува „не знам“ исто како погрешен одговор, па погодувањето секогаш победува — до степен модел што греши три четвртини од времето да може да изгледа подобро од поискрениот модел што се воздржува. Предлогот на авторите не е уште еден тест за халуцинации, туку „open rubrics“: правилата за бодирање се наведуваат во самото прашање. Во case study со четири frontier модели, тоа го превртува поттикот така што метод што ги намалува халуцинациите се наградува наместо да се казнува. Ова е теоретски труд плус преглед со мал, експлицитно неконтролиран експеримент, рецензиран во *Nature*; поправката е ветувачка, но сè уште не е покажано дека работи широко и во голем обем, а халуцинациите се претставуваат како ниту мистични ниту строго неизбежни.

Проверка без претерување

Што покажува трудот: математичка долна граница според која одредена халуцинација е присилена при pretraining — најмалку singleton-стапката за факти без образец; преглед што открива дека повеќето водечки benchmark-и не му даваат кредит на „не знам“; и case study со четири модели каде наведувањето на правилата за бодирање во prompt-от, „open rubrics“, прави метод за намалување халуцинации да победи таму каде обичната accuracy го казнувала.

Што е веројатно, но не е докажано: дека open rubrics, ако се вградат во главните benchmark-и, значајно би ги намалиле халуцинациите во применети модели. Поддржувачкиот експеримент е мал и експлицитно неконтролиран.

Што не покажува: дека халуцинациите се неизбежни — трудот тврди спротивно;

дека бодирањето е единствена причина; дека халуцинациите можат целосно да се елиминираат; дека case study ги рангира четирите модели меѓусебно.

Главни ограничувања: намерно поедноставен модел „точно/погрешно/не знам“, што авторите го нарекуваат „false trichotomy“; мал куриран преглед од десет евалуации; неконтролиран case study со четири модели, една мерка, еден тест и default поставки; и аргумент предводен од OpenAI за тоа како полето треба да ги оценува моделите.

Колку доверба треба да има општ читател? Висока дека халуцинациите не се ниту мистериозни ниту строго неизбежни и дека главните benchmark-и во моментов го наградуваат погодувањето. Средна дека предложената поправка помага — сега има реален proof of concept, но не и демонстрација дека работи широко и во голем обем.

Извори

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