The Clean Paper

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

Защо езиковите модели халюцинират — и защо начинът, по който ги оценяваме, поддържа проблема

Уверените неверни твърдения, които наричаме „халюцинации“, не са мистериозен дефект: част от тях са статистически страничен продукт на обучението и продължават да се появяват, защото масовите benchmarks награждават уверено предположение повече от честно „не знам“. Case study върху четири frontier модела показва, че изписването на правилата за оценяване в самия prompt („open rubrics“) обръща този стимул.

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

Защо моделите гадаят и защо ние ги научихме да го правят

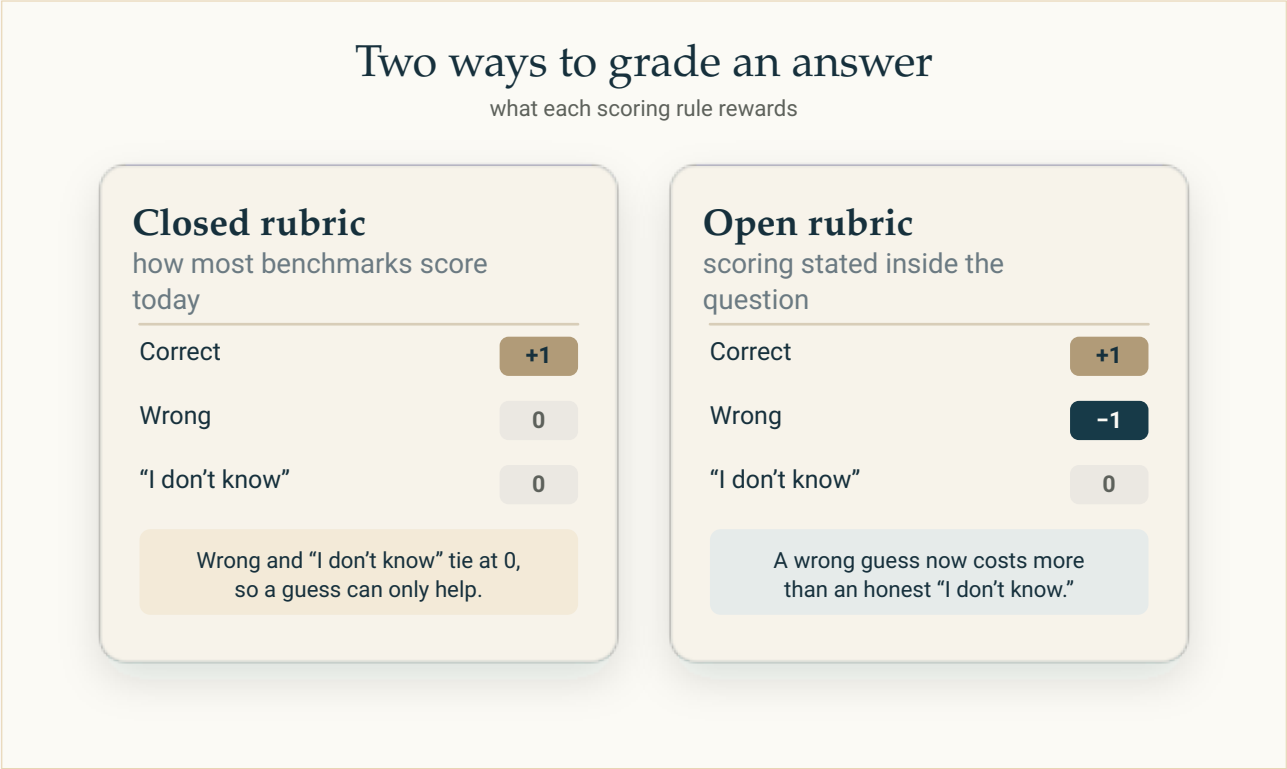
Попитайте голям езиков модел за рождения ден на непознат човек и той може да отговори „7 март“ с увереността на някой, който го чете от карта — и да греши три пъти поред, всеки път с различна дата. Авторите дават точно такива примери: водещи модели получават прост фактологичен въпрос — рожден ден на човек или значението на рядко съкращение — и всеки уверено измисля различен отговор, без нито един да е правилен. Индустрията нарича това *халюцинация*, което звучи като дефект на възприятието. Първата стъпка на статията е да махне мистерията.

Да започнем от начина, по който се изгражда моделът. В първия и най-голям етап на обучение той научава, по същество, как изглежда плавният език, като прочита огромно количество текст. Сега вземете факт, зад който няма закономерност — рождения ден на конкретен човек. Ако тази дата се е появила в обучителния текст веднъж или никога, няма закономерност, за която обучаващият се модел да се хване: от гледна точка на системата отговорът е произволен. Авторите формализират това, като заемат стара идея (на Alan Turing, от съвсем друг проблем): ако един от пет рождени дни се появява в данните само веднъж, моделът би трябвало да греши поне при един от пет — не защото е счупен, а защото никога не е имало достатъчно информация, която да научи. (По същата логика моделите почти никога не бъркат столица на държава: тези факти се появяват постоянно.) Авторите внимателно аргументират, че разграничаването на вярно твърдение от правдоподобно невярно само по себе си е трудна задача и че генерирането *само* на верни твърдения е поне толкова трудно. Има вграден долен праг на грешките.

> **Идеята отдолу: как да преброиш онова, което още не си виждал**
>
> Това се опира на действително умна идея — по-стара от езиковите модели и заслужаваща да бъде разбрана сама по себе си.
> **Започнете с торба цветни топчета.** Не знаете колко цвята има вътре. Изтегляте 100, едно по едно, и ги броите: червени 40, сини 25, зелени 15, жълти 5, лилави 3, оранжеви 2 — и после **десет различни цвята, които се появяват точно по веднъж.**
> Сега въпросът, пред който Turing е бил изправен при съвсем различен проблем: *каква е вероятността следващото топче да е цвят, който изобщо не сте виждали?* Не можете да преброите нещо, което никога не сте изтеглили — но **можете** да преброите цветовете, видени точно веднъж, „единични наблюдения“ (*singletons*). Трикът, наречен **оценка на Гуд–Тюринг**, е, че делът от изтеглянията, които са единични наблюдения, оценява вероятността, скрита в цветовете, които още не сте виждали. Десет от сто изтегляния са цветове, появили се само веднъж, значи шансът следващото топче да е напълно нов цвят е около **10 / 100 = 10%.**
> Тези цветове, видени веднъж, не са *грешки*. Те са измерване на собственото ви незнание: много цветове, появили се по веднъж, са начинът извадката да ви

> каже, че светът съдържа още неща, които просто не сте изтеглили.
> **Сега заменете цветовете с рождени дни,** а торбата — с обучителния текст на модела. Да предположим, че сред рождените дни, които е видял, **един от пет се появява точно веднъж.** Същият трик: около една пета от вероятността се намира в рождени дни, които моделът на практика никога не е виждал — а рожденият ден няма закономерност, върху която да стъпи (не можете *да изчислите* нечий рожден ден). Така дата, видяна веднъж или никога, е монета, която моделът не може да претегли, и той ще греши при приблизително **един от пет** случая. Никаква изобретателност не поправя това: просто е нямало какво да се научи.
> Това е целият аргумент в миниатюра: делът на единичните наблюдения измерва колко от света е ненаучим *от тези данни* и това се превръща в **долен праг** за грешките. И затова модел почти никога не бърка столица — *Париж* се появява постоянно, делът на единичните наблюдения е почти нула и има много материал за научаване.

Това обяснява откъде идват халюцинациите. Не обяснява защо *оцеляват* — защо след всички по-късни етапи на обучение, предназначени да направят моделите полезни и честни, те продължават да блъфират вместо да признаят съмнение. Тук аналогията на статията е почти неприятно точна. Представете си ученик на изпит, който не знае отговора. Ако празният отговор носи нула, а предположението може да донесе една точка, стратегията за максимален резултат е да гадае — уверено, конкретно, никога „не съм сигурен“. Учениците научават това. Оказва се, и моделите — защото ги оценяваме по същия начин. Авторите преглеждат тестовете за оценяване, по които областта действително се състезава, и класациите, за чието изкачване моделите се настройват, и установяват, че почти всички дават на „не знам“ точно същия резултат като на грешен отговор: нула. При такова правило модел, който винаги гадае, ще победи иначе идентичен модел, който честно отбелязва несигурността си. В доста буквален смисъл ние ги оценяваме така, че да халюцинират.



Тестовете за оценяване могат да направят гадаенето рационално. При оценяването, използвано от повечето такива тестове (вляво), грешен отговор и честно „не знам“ и двете носят нула — така предположението може само да помогне. Ако правилата са написани в самия въпрос и има наказание за грешка (вдясно), въздържането при несигурност може да стане по-добрият ход. Това променя какво награждава тестът; само по себе си не решава халюцинациите.

Original diagram — The Clean Paper CC BY 4.0

Това е частта, която си струва да остане, защото върви срещу обичайното заглавие. Халюцинацията често се продава като неизбежна, почти мистична граница на технологията. Статията оспорва и двете. Долният праг от предварителното обучение не е мистерия — това е обикновена статистическа грешка, позната на машинното обучение от десетилетия. А устойчивото ѝ присъствие не е неизбежно — отчасти е стимул, който сме създали и можем да променим. Система, която просто отказва да отговори, когато не е сигурна, изобщо не би халюцинирала; причината внедрените модели да не се държат така е, че нашите таблици с резултати наказват отказа.

Какво са направили авторите

Статията има три части. Първо, математически аргумент, че известна халюцинация е статистически принудена по време на предварителното обучение, като показва, че „генерирай само валиден текст“ е поне толкова трудно, колкото двоична задача за класификация „валидно ли е това твърдение?“. Второ, аргумент — подкрепен с обзор на десет влиятелни теста за оценяване — че масовите метрики, основани на точността награждават гадаенето повече от въздържането. Трето, предложена

поправка и примерен анализ, който я тества: оценки с **открити правила**, при които правилата за оценяване са написани вътре в самия въпрос (например „правилен отговор носи 1, грешен −1, затова се въздържай, ако си под 50% сигурност“), така че моделът да може да разбере кога честността се награждава. Те тестват това върху четири водещи модела — Gemini 3 Pro на Google, GPT-5 на OpenAI, Grok 4 на xAI и Claude Opus 4.5 на Anthropic — чрез 4326-те фактологични въпроса на SimpleQA. Изрично посочват, че примерният анализ е илюстративен, „не контролирана оценка между модели“ (настройки по подразбиране, без донастройване, без нормализиране спрямо разходите).

Какво са установили

- **Предварителното обучение налага известна грешка.** Честотата, с която модел излъчва уверени неверни твърдения, има долна граница, определяна от (приблизително два пъти) грешката на най-добрия класификатор „валидно ли е твърдението?“, построен от него. За факти без научима закономерност този праг е поне *дял на единичните срещания* — делът от факти, които се появяват точно веднъж в обучението. Известна халюцинация е неизбежна дори при идеално чисти данни.
- **Оценяването награждава гадаенето — конкретно.** При обикновено оценяване тип правилно/грешно оптималната стратегия е никога да не се въздържаш, а обзорът на авторите установява, че огромното мнозинство популярни тестове за оценяване оценяват „не знам“ просто като грешно. Ярък пример от собствените им модели: в SimpleQA простата точност леко *предпочита* o4-mini на OpenAI — който отговаря почти на всичко и греши в повече от три четвърти от случаите — пред GPT-5-mini, който прави много по-малко грешки, защото се въздържа, когато не е сигурен. По-безразсъдният модел изглежда по-добре на класацията.
- **Откритите правила обръщат стимула (в техния примерен анализ).** Те тестват прост метод за ограничаване на халюцинациите (моделът отговаря два пъти и се въздържа, ако двата отговора не съвпадат). При стандартно измерване на точността методът намалява грешките, *но намалява и измерената точност* — така метриката обезкуражава внедряването му. При открити правила за оценяване същият метод излиза по-добре за всичките четири модела при диапазон от наказания; и GPT-5-mini — който простата точност е наказвала за въздържане при несигурност — излиза пред o4-mini, след като правилото за оценяване е заявено открито (n = 4326 въпроса на модел).

Какво вероятно означава това

Намаляването на халюцинациите до голяма степен не е въпрос на изобретяване на още тестове, специално насочени към халюцинациите. То е въпрос на промяна в начина, по който масовите тестове за оценяване оценяват несигурността, така

че признаването „не знам“ вече да не бъде наказвано. Докато класацията не се промени, намаляването на халюцинациите ще продължи да струва точки за точност и следователно ще бъде обезкуражавано — затова авторите наричат проблема „социотехнически“: отчасти по-добра метрика, отчасти убеждаване на влиятелните класации да я приемат.

Какво *не* доказва това

- То **не** показва, че открити правила за оценяване решават халюцинацията в реални условия. Подкрепящият експеримент е малък, умишлено **неконтролиран** примерен анализ — четири модела с настройки по подразбиране, един избран метод за ограничаване, един тест с фактологични въпроси и отговори — предназначен да демонстрира *обръщането на стимула*, не да класира модели или да докаже обща ефективност.
- То **не** твърди, че оценяването е *единствената* причина. Грешки в обучителните данни, действително трудни проблеми и непознати инструкции остават отделни източници.
- То **не** подкрепя популярната теза, че халюцинациите са неизбежни. Авторите твърдят обратното: система, която отговаря само на проверими въпроси и иначе казва „не знам“, никога не би халюцинирала.
- То **не** премахва долния праг от предварителното обучение — обяснява го и го ограничава, а тази граница се отнася до уверени фактологични грешки, не до цялото поведение на модела.
- То **не** показва, че открити правила за оценяване са *достатъчни* сами по себе си. Те променят какво награждава оценката; не заместват извличането на информация, използването на инструменти или по-добре калибрирани модели.

Колко силни са доказателствата?

- Ядрото е **математика** — формални долни граници, не измервания. Като теоретичен аргумент е валидно в собствените си допускания.
- То се опира на **умишлено опростени модели** на проблема; самите автори посочват „фалшива тройна схема“ при третирането на всеки отговор като правилен, неправилен или „не знам“, и идеализираната постановка на „произволни факти“, използван за най-чистата граница.
- Обзорът на тестовете за оценяване е **малка, подбрана извадка** — десет влиятелни оценки, не изчерпателен одит.
- Примерният анализ е **реално, но ограничено**: четири водещи модела, един метод за ограничаване, само SimpleQA, настройки по подразбиране, изрично „не е контролирана оценка“. Това е доказателство на концепцията за аргумента за стимулите, не резултат от тест за оценяване.

- Струва си да се назове гледната точка: трима от четиримата автори са или са били служители на **OpenAI**, а статията твърди, че областта трябва да промени начина, по който оценява моделите. Това е добре аргументирана позиция на заинтересована страна, не неутрален външен обзор — трябва да се претегли, не да се отхвърли. (В нейна полза е, че статията насочва критиката толкова охотно към собствените модели o4-mini и GPT-5-mini, колкото към чуждите.)

Защо това има значение

То преформулира силно преувеличаван проблем. „Халюцинация“ обикновено се продава или като призрачен дефект, или като неподвижна стена; тази статия я прави обикновена и отчасти самопричинена — статистически праг, който можем да разберем, върху стимул, който сами сме избрали. По-широкият урок е по-тих и по-полезен: бъдещият напредък в надеждността може да зависи толкова от *това, което измерваме*, колкото от това, което изграждаме.

Накратко

Уверените неверни отговори на езиковите модели идват от две места. Първото е статистическо: когато фактът няма закономерност, която да бъде научена, модел, обучен да имитира език, понякога ще греши, а този праг може да бъде оценен (например от броя факти, които се появяват само веднъж в обучението). Второто са стимулите: почти всеки тест за оценяване, по който моделите се класират, оценява „не знам“ също като грешен отговор, така че гадаенето винаги печели — до степен модел, който греши в три четвърти от случаите, да може да изпревари по-честен модел, който се въздържа. Предложението на авторите не е още един тест за халюцинации, а „открити правила за оценяване“: правилата за оценяване да се заявяват вътре във въпроса. В примерен анализ върху четири водещи модела това обръща стимула, така че метод за намаляване на халюцинациите се награждава, вместо да бъде наказван. Това е теория плюс обзор с малък, изрично неконтролиран експеримент, рецензирано в *Nature*; поправката е обещаваща, но още не е показано, че работи в голям мащаб, а аргументът е, че халюцинациите не са нито мистериозни, нито строго неизбежни.

Проверка без украса

Какво показва статията: Математическа долна граница, при която известна халюцинация е принудена по време на предварителното обучение (поне „дял на единичните срещания“ при факти без закономерност); обзор, според който повечето водещи тестове за оценяване не дават кредит на „не знам“; и примерен анализ с четири

модела, при който изписването на правилото за оценяване в самата инструкция („открити правила за оценяване“) кара метод за намаляване на халюцинациите да печели там, където обикновената точност го е наказвала.

Какво е правдоподобно, но недоказано: Че открити правила за оценяване, ако бъдат включени в масовите тестове за оценяване, биха намалили съществено халюцинациите в внедрени модели. Подкрепящият експеримент е малък и изрично неконтролиран.

Какво не показва: Че халюцинациите са неизбежни (аргументът е обратният); че оценяването е единствената причина; че халюцинацията може да бъде премахната напълно; че примерният анализ класира четирите модела един спрямо друг.

Основни ограничения: Умишлено опростен модел правилно/грешно/„не знам“ (авторите го наричат „фалшива тройна схема“); малък подбран обзор на тестове за оценяване (десет оценки); неконтролиран примерен анализ (четири модела, един метод за ограничаване, един тест, настройки по подразбиране); и аргумент, воден от OpenAI за това как областта трябва да оценява модели.

Колко доверие трябва да има общият читател? Високо, че халюцинацията не е нито мистериозна, нито строго неизбежна и че масовите тестове за оценяване в момента награждават гадаенето. Средно, че предложената поправка помага — вече има реален доказателство на концепцията, но още няма демонстрация, че работи широко и в мащаб.

Източници

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