The Clean Paper

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

Чому мовні моделі галюцинують — і чому наше оцінювання допомагає цьому тривати

Впевнені хибні твердження, які ми називаємо «галюцинаціями», не є загадковим збоєм: частина з них — статистичний побічний ефект навчання, а зберігаються вони ще й тому, що головні бенчмарки винагороджують упевнену здогадку більше за чесне «я не знаю». Case study на чотирьох frontier-моделях показує, що явне формулювання правил оцінювання в самому запиті — «open rubrics» — перевертає цей 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

Чому моделі вгадують — і чому ми самі їх цьому навчили

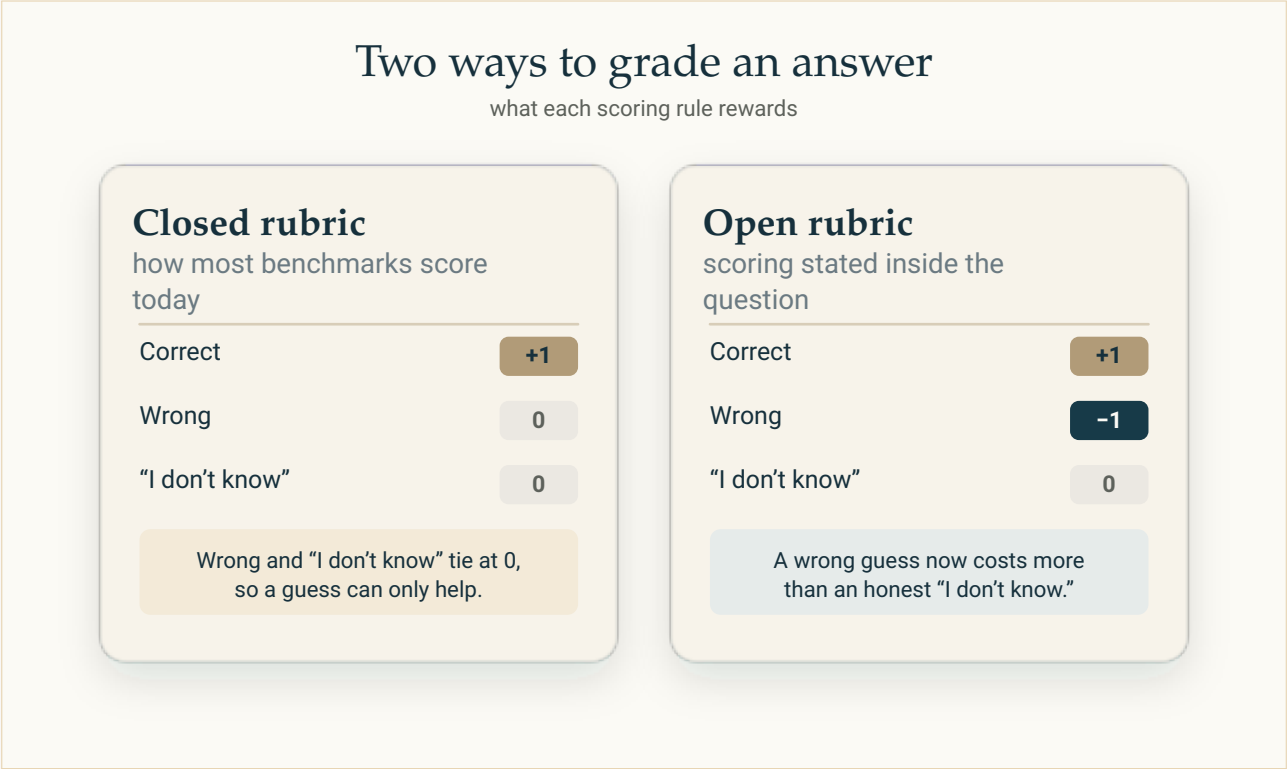
Запитайте велику мовну модель про день народження незнайомої людини — і вона може відповісти «7 березня» зі спокійною впевненістю того, хто читає дату з картки, а потім тричі поспіль помилитися, щоразу назвавши іншу дату. Автори наводять саме такі приклади: провідні моделі отримують просте фактичне запитання — день народження людини або розшифрування маловідомої абревіатури — і кожна впевнено вигадує свою відповідь, причому жодна не правильна. У галузі це називають *галюцинацією*, ніби йдеться про збій сприйняття. Перший крок статті — прибрати з цього явища містичність.

Почнімо з того, як будують модель. На першому й найбільшому етапі навчання вона, по суті, вчиться тому, як виглядає природна мова, читаючи величезний масив тексту. Тепер візьмімо факт без закономірності — конкретний день народження певної людини. Якщо ця дата трапилася в навчальному тексті один раз або взагалі не трапилася, моделі, що вчиться на патернах, нема за що зачепитися: з її точки зору відповідь довільна. Автори роблять цю ідею точною, використовуючи старий підхід Алана Тюрінга з іншої задачі: якщо кожен п'ятий день народження зустрічається в даних лише один раз, слід очікувати, що модель помилятиметься щонайменше приблизно на кожному п'ятому — не тому, що вона «зламана», а тому, що в даних не було закономірності, яку можна вивчити. (За тією самою логікою моделі майже ніколи не помиляються зі столицями країн: ці факти повторюються постійно.) Автори обережно доводять, що відрізнити істинне твердження від правдоподібного хибного — саме по собі складна задача, а генерувати *лише* істинні твердження щонайменше не простіше. Отже, певна нижня межа помилок закладена статистично.

> **Ідея під цим результатом: як порахувати те, чого ви ще не бачили**
>
> Тут використано справді красиву ідею — старшу за мовні моделі й варту окремого знайомства.
> **Почнімо з мішка кольорових кульок.** Ви не знаєте, скільки в ньому кольорів. Витягуєте 100 кульок по одній і рахуєте: червоних 40, синіх 25, зелених 15, жовтих 5, фіолетових 3, помаранчевих 2 — а потім **десять різних кольорів, кожен із яких трапився рівно один раз**.
> Тепер питання, з яким Тюрінг насправді зіткнувся в зовсім іншій задачі: *яка ймовірність, що наступна кулька матиме колір, якого ми ще взагалі не бачили?* Порахувати те, чого ви ніколи не витягували, неможливо — але **можна** порахувати кольори, які трапилися рівно один раз, тобто «singletons». Трюк, відомий як **оцінювання Гуда–Тюрінга (Good–Turing estimation)**, полягає в тому, що частка одноразових спостережень оцінює ймовірнісну масу, яка ще ховається в невідомих категоріях. Десять зі ста витягнутих кульок належали до кольорів, що трапилися лише раз, тож імовірність, що наступний колір буде абсолютно новим, становить приблизно **10 / 100 = 10%.**
> Ці одноразові кольори — не *помилки*. Вони є вимірюванням вашого власного

> незнання: велика кількість кольорів, що з'явилися один раз, — це спосіб, яким вибірка повідомляє, що у світі є ще багато того, чого ви просто не витягнули.
> **Тепер замінімо кольори на дні народження,** а мішок — на навчальний текст моделі. Припустімо, серед побачених днів народження **кожен п'ятий трапляється рівно один раз.** Той самий трюк: приблизно п'ята частина ймовірнісної маси припадає на дні народження, яких модель фактично ніколи не бачила, а в дня народження немає закономірності, на яку можна спертися — його не можна *вирахувати*. Тож дата, побачена один раз або ніколи, — це монета, вагу якої модель не знає, і вона помилятиметься приблизно в **одному випадку з п'яти**. Жодна «розумність» не виправить відсутність інформації: вчитися було нема з чого.
> У мініатюрі це весь аргумент: частка singletons вимірює, скільки світу *неможливо вивчити з цих даних*, і це стає **нижньою межею** помилок. Саме тому модель майже ніколи не промахується зі столицею: *Paris* зустрічається постійно, його singleton rate близький до нуля, отже інформації для навчання багато.

Це пояснює, звідки беруться галюцинації. Але не пояснює, чому вони *виживають* — чому після всіх наступних етапів навчання, покликаних зробити модель корисною та чесною, вона все одно блефує замість того, щоб визнати невпевненість. Тут аналогія авторів майже незручно точна. Уявіть студента на іспиті, який не знає відповіді. Якщо порожня відповідь дає нуль, а здогадка може дати один бал, стратегія, що максимізує оцінку, — вгадувати: впевнено, конкретно, ніколи не писати «я не знаю». Студенти цьому вчаться. Як виявляється, моделі теж — бо ми оцінюємо їх так само. Автори переглянули бенчмарки, на яких реально змагається галузь, рейтинги, під які моделі оптимізують, і виявили: майже всі дають «я не знаю» рівно стільки ж балів, скільки неправильній відповіді — нуль. За такого правила модель, що завжди вгадує, випередить ідентичну модель, яка чесно позначає невпевненість. У досить буквальному сенсі ми самі оцінюванням підштовхуємо їх до цього.



Бенчмарки можуть робити вгадування раціональною стратегією. За схемою оцінювання, яку використовує більшість тестів (ліворуч), неправильна відповідь і чесне «я не знаю» обидва дають нуль — отже, здогадка може лише допомогти. Якщо правила явно записати в самому запитанні й штрафувати за помилку (праворуч), відмова від відповіді за невпевненості може стати кращим вибором. Це змінює те, що винагороджує тест; саме по собі воно не розв’язує проблему галюцинацій.

Original diagram — The Clean Paper CC BY 4.0

Саме цю частину варто запам’ятати, бо вона суперечить звичному заголовку. Галюцинації часто подають як неминучу, майже містичну межу технології. Стаття заперечує обидві частини. Нижня межа на етапі pretraining не є таємницею — це звичайна статистична помилка, яку машинне навчання розуміє десятиліттями. А її збереження не неминуче: частково це стимул, який ми самі створили й можемо змінити. Система, що просто відмовляється відповідати, коли не впевнена, взагалі не галюцинувала б; розгорнуті моделі не поводяться так тому, що наші таблиці результатів карають відмову.

Що зробили автори

Стаття має три частини. Перша — математичний аргумент, що певна частка галюцинацій статистично неминуча під час pretraining: автори показують, що задача «генерувати лише коректний текст» щонайменше така сама складна, як бінарна класифікація «це твердження коректне чи ні?». Друга — аргумент, підкріплений оглядом десяти впливових бенчмарків, що звичайні метрики точності винагороджують вгадування більше за утримання від відповіді. Третя — запропоноване виправлення та case study: **open-rubric** оцінювання, де правило підрахунку балів прямо записано в запитанні

(наприклад, «правильна відповідь дає 1 бал, неправильна −1, тому утримайтеся, якщо ваша впевненість менша за 50%»), щоб модель знала, коли чесність винагороджується. Це перевіряють на чотирьох frontier-моделях — Google's Gemini 3 Pro, OpenAI's GPT-5, xAI's Grok 4 та Anthropic's Claude Opus 4.5 — використовуючи 4 326 фактичних запитань SimpleQA. Автори прямо пишуть, що case study лише ілюстративний і «не є контрольованим порівнянням моделей» — стандартні налаштування, без тюнінгу й нормалізації вартості.

Що вони знайшли

- **Pretraining змушує мати певну частку помилок.** Частота, з якою модель видає впевнені хибні твердження, має нижню межу, приблизно пов'язану з подвоєною частотою помилки найкращого класифікатора «чи коректне це твердження?», побудованого з неї. Для фактів без закономірності, яку можна вивчити, ця межа щонайменше дорівнює *singleton rate* — частці фактів, які зустрічаються в навчанні рівно один раз. Певні галюцинації неминучі навіть у абсолютно чистих даних.
- **Оцінювання конкретно винагороджує вгадування.** За звичайної схеми «правильно/неправ оптимальна стратегія — ніколи не утримуватися від відповіді, а огляд авторів показує, що переважна більшість популярних бенчмарків зараховує «я не знаю» просто як помилку. Яскравий приклад із їхніх власних моделей: на SimpleQA сира accuracy трохи *віддає перевагу* OpenAI o4-mini, яка відповідає майже на все й помиляється більш ніж у трьох чвертях випадків, над GPT-5-mini, яка робить набагато менше помилок, бо утримується, коли не впевнена. Безрозсудніша модель виглядає кращою в таблиці.
- **Open rubrics перевертають стимул — у їхньому case study.** Автори тестують простий спосіб зменшення галюцинацій: попросити модель відповісти двічі й утриматися, якщо дві відповіді не збігаються. За стандартної accuracy метод зменшує помилки, *але водночас зменшує accuracy*, тож метрика фактично карає його використання. За open rubrics той самий метод виходить кращим для всіх чотирьох моделей у широкому діапазоні штрафів; а GPT-5-mini, яку сира accuracy карала за відмову від відповіді за невпевненості, випереджає o4-mini, коли схема балів сформульована відкрито (n = 4 326 запитань на модель).

Що це, ймовірно, означає

Зменшення галюцинацій переважно не потребує вигадувати ще більше спеціальних тестів на галюцинації. Потрібно змінити те, як головні бенчмарки оцінюють невпевненість, щоб визнання «я не знаю» перестало каратися. Доки таблиця результатів не зміниться, зменшення галюцинацій і далі коштуватиме моделям балів accuracy й тому буде невигідним — саме тому автори називають проблему «соціотехнічною»: потрібна і краща метрика, і готовність впливових рейтингів її прийняти.

Чого ця робота *не* доводить

- Вона **не** показує, що open rubrics виправляють галюцинації в реальному використанні. Підтримувальний експеримент — маленький і навмисно **неконтрольований** case study: чотири моделі зі стандартними налаштуваннями, один вибраний спосіб зменшення галюцинацій, один тест фактичних запитань. Його мета — показати *переворот стимулу*, а не ранжувати моделі чи довести загальну ефективність.
- Вона **не** стверджує, що оцінювання — *єдина* причина. Помилки в навчальних даних, справді складні задачі й незнайомі запити залишаються окремими джерелами.
- Вона **не** підтримує популярне твердження, що галюцинації неминучі. Автори стверджують протилежне: система, яка відповідає лише на перевірювані запитання й інакше каже «я не знаю», ніколи не галюцинувала б.
- Вона **не** прибирає нижню межу pretraining — вона її пояснює й обмежує, а сама межа стосується впевнених фактичних помилок, не всієї поведінки моделі.
- Вона **не** показує, що open rubrics *достатньо* самі по собі. Вони змінюють те, що винагороджує оцінювання; вони не замінюють retrieval, використання інструментів чи краще калібровані моделі.

Наскільки переконливі докази?

- Ядро роботи — **математика**: формальні нижні межі, а не вимірювання. Як теоретичний аргумент воно послідовне в межах власних припущень.
- Воно спирається на **навмисно спрощені моделі** проблеми. Самі автори позначають «хибну трихотомію» — поділ кожної відповіді лише на правильну, неправильну або «я не знаю» — і ідеалізоване середовище «довільних фактів», використане для найчіткішої межі.
- Огляд бенчмарків — **маленька, вручну відібрана вибірка**: десять впливових оцінювань, а не повний аудит.
- Case study **реальний, але обмежений**: чотири frontier-моделі, один mitigation, лише SimpleQA, стандартні налаштування, і прямо сказано «не контрольоване порівняння». Це proof of concept для аргументу про стимули, а не новий рейтинг моделей.
- Варто назвати й позицію авторів: троє з чотирьох працюють або працювали в **OpenAI**, а стаття аргументує, як галузі слід змінити оцінювання моделей. Це добре аргументована позиція зацікавленої сторони, а не нейтральний зовнішній огляд — її слід зважувати, а не відкидати. (До честі статті, вона спрямовує критику на власні моделі o4-mini та GPT-5-mini так само охоче, як на інші.)

Чому це важливо

Робота переосмислює надзвичайно розкручену проблему. «Галюцинацію» часто продають або як моторошний дефект, або як нерухому стіну; ця стаття робить її звичайнішою й частково самоствореною — статистична нижня межа, яку ми можемо зрозуміти, плюс стимул, який ми самі вибрали. Ширший урок тихіший і корисніший: подальший прогрес у надійності може залежати не менше від *того, що ми вимірюємо*, ніж від того, що ми будуємо.

Короткий підсумок

Впевнені хибні відповіді мовних моделей мають два джерела. Перше статистичне: коли факт не має закономірності, яку можна вивчити, модель, навчена імітувати мову, іноді помилятиметься, і цю нижню межу можна оцінити — наприклад, за часткою фактів, що трапляються в навчанні лише один раз. Друге джерело — стимули: майже кожен бенчмарк, за яким ранжують моделі, оцінює «я не знаю» так само, як неправильну відповідь, тому вгадування завжди вигідніше — аж до ситуації, коли модель, що помиляється в трьох чвертях випадків, може випередити чеснішу модель, яка утримується від відповіді. Пропозиція авторів — не ще один тест на галюцинації, а **open rubrics**: записувати правило оцінювання прямо в запитанні. У case study на чотирьох frontier-моделях це перевертає стимул, і метод зменшення галюцинацій винагороджується, а не карається. Це теоретична робота плюс огляд із невеликим, явно неконтрольованим експериментом, рецензована в *Nature*. Виправлення перспективне, але ще не показане у великому масштабі; самі галюцинації тут трактуються як не містичні й не строго неминучі.

Твереза оцінка

Що показує стаття: Математичну нижню межу, за якої деякі галюцинації виникають уже під час pretraining — для фактів без патернів щонайменше на рівні singleton rate; огляд, за яким більшість провідних бенчмарків не дають жодного бонусу за «я не знаю»; і case study на чотирьох моделях, де явне формулювання правил оцінювання в запиті — open rubrics — робить метод зменшення галюцинацій вигідним там, де проста accuracy його карала.

Що правдоподібно, але не доведено: Що впровадження open rubrics у головні бенчмарки суттєво зменшить галюцинації в розгорнутих моделях. Підтримувальний експеримент невеликий і прямо названий неконтрольованим.

Чого вона не показує: Що галюцинації неминучі — робота стверджує протилежне; що оцінювання є єдиною причиною; що галюцинації можна повністю усунути; що case

study ранжує чотири моделі між собою.

Головні обмеження: Навмисно спрощена модель «правильно/неправильно/я не знаю», яку самі автори називають «хибною трихотомією»; невеликий вручну відібраний огляд бенчмарків — десять оцінювань; неконтрольований case study — чотири моделі, один mitigation, один тест, стандартні налаштування; і аргумент, значною мірою підготовлений дослідниками OpenAI, про те, як галузі слід оцінювати моделі.

Якої впевненості заслуговує висновок для загального читача? Високої, що галюцинації не є ні містичними, ні строго неминучими, і що головні бенчмарки нині винагороджують вгадування. Помірної, що запропоноване виправлення допоможе: тепер є справжній proof of concept, але ще немає демонстрації широкої ефективності у великому масштабі.

Джерела

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