The Clean Paper

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

Dlaczego modele językowe halucynują — i dlaczego nasz sposób oceniania utrwala ten problem

Pewnie podawane fałsze, które nazywamy „halucynacjami", nie są tajemniczą usterką: część stanowi statystyczny produkt uboczny treningu, a problem utrzymuje się, ponieważ główne benchmarki nagradzają pewne siebie zgadywanie bardziej niż uczciwe „nie wiem". Studium czterech czołowych modeli pokazuje, że podanie zasad punktacji w poleceniu — zastosowanie „jawnych kryteriów" — odwraca tę zachętę.

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

Dlaczego modele zgadują — i dlaczego je tego nauczyliśmy

Zapytaj duży model językowy o datę urodzenia nieznanej osoby, a może odpowiedzieć „7 marca" z pewnością kogoś, kto odczytuje ją z dokumentu — i trzy razy z rzędu podać trzy różne, błędne daty. Autorzy pokazują dokładnie takie przykłady: czołowe modele dostają proste pytanie o fakt — urodziny danej osoby albo rozwinięcie mało znanego skrótu — i każdy z przekonaniem wymyśla inną odpowiedź, wszystkie niepoprawne. Branża nazywa to *halucynacją*, jak gdyby chodziło o zaburzenie percepcji. Pierwszym krokiem pracy jest odarcie tego z tajemnicy.

Zacznijmy od sposobu budowy modelu. W pierwszym i największym etapie treningu uczy się on w praktyce, jak wygląda płynny język, czytając olbrzymią ilość tekstu. Weźmy teraz fakt, za którym nie kryje się żaden wzorzec — datę urodzenia konkretnej osoby. Jeśli pojawiła się w danych treningowych raz albo wcale, model uczący się wzorców nie ma czego uchwycić: z jego punktu widzenia odpowiedź jest arbitralna. Autorzy precyzują to, zapożyczając dawny pomysł Alana Turinga z innego problemu: jeśli co piąta data urodzenia występuje w danych tylko raz, należy oczekiwać, że model pomyli co najmniej jedną na pięć — nie dlatego, że jest zepsuty, lecz dlatego, że w danych nie było niczego, czego mógłby się nauczyć. Z tej samej przyczyny modele niemal nigdy nie mylą stolicy państwa: takie informacje pojawiają się bez przerwy. Autorzy ostrożnie argumentują, że odróżnienie prawdziwego zdania od wiarygodnie brzmiącego fałszu samo w sobie jest trudnym zadaniem, a generowanie *wyłącznie* prawdziwych zdań jest co najmniej równie trudne. Istnieje więc nieusuwalny dolny poziom błędów.

Idea u podstaw: jak policzyć to, czego jeszcze nie widzieliśmy

Opiera się to na naprawdę pomysłowej idei — starszej niż modele językowe i wartej dokładnego poznania.

Zacznijmy od worka kolorowych kul. Nie wiesz, ile jest w nim kolorów. Losujesz kolejno 100 kul i zapisujesz wyniki: 40 czerwonych, 25 niebieskich, 15 zielonych, 5 żółtych, 3 fioletowe, 2 pomarańczowe — a następnie **dziesięć różnych kolorów, z których każdy pojawia się dokładnie raz.**

Turing, pracując nad zupełnie innym problemem, pytał: *jakie jest prawdopodobieństwo, że następna kula będzie miała kolor, którego dotąd w ogóle nie widzieliśmy?* Nie da się policzyć kolorów, których nigdy nie wylosowano — można jednak policzyć kolory widziane dokładnie raz, czyli „singletony". Sztuczka zwana **estymacją Gooda–Turinga** polega na tym, że udział pojedynczych wystąpień w próbie szacuje prawdopodobieństwo ukryte w niewidzianych jeszcze kolorach. Dziesięć ze stu losowań dało kolory jednorazowe, więc szansa, że następna kula będzie miała zupełnie nowy kolor, wynosi około **10 / 100 = 10%.**

Kolory widziane raz nie są *błędami*. Są miarą naszej niewiedzy: duża liczba pojedynczych wystąpień to sposób, w jaki próba mówi nam, że świat zawiera więcej możliwości, których jeszcze nie wylosowaliśmy.

Zastąpmy teraz kolory datami urodzenia, a worek — tekstem treningowym modelu. Załóżmy, że wśród dat urodzenia widzianych przez model **jedna na pięć pojawiła się dokładnie raz.** Ta sama sztuczka mówi, że około jednej piątej prawdopodobieństwa przypada na daty, których model w praktyce nigdy nie poznał — a dla daty urodzenia nie ma wzorca zastępczego, bo

> nie można jej *wywnioskować*. Data widziana raz albo niewidziana wcale jest więc rzutem monetą, której model nie potrafi odpowiednio obciążyć, i w przybliżeniu **jedną na pięć** takich odpowiedzi poda błędnie. Żaden spryt tego nie naprawi: w danych po prostu nie było materiału do nauki.
> To cały argument w miniaturze: odsetek singletonów mierzy, jakiej części świata nie da się nauczyć *z tych danych*, a to wyznacza **dolną granicę** błędów. Wyjaśnia też, dlaczego model prawie nigdy nie myli stolicy — *Paryż* występuje bez przerwy, odsetek jego pojedynczych wystąpień jest bliski zeru, więc materiału do nauki jest mnóstwo.

To wyjaśnia, skąd biorą się halucynacje. Nie mówi jednak, dlaczego *przetrwały* — dlaczego po późniejszych etapach treningu, które mają uczynić modele pomocnymi i uczciwymi, nadal blefują zamiast przyznać się do wątpliwości. Analogia autorów jest tu niemal boleśnie trafna. Wyobraźmy sobie ucznia, który nie zna odpowiedzi na egzaminie. Jeśli puste miejsce daje zero punktów, a zgadywanie może dać jeden, strategią maksymalizującą wynik jest zgadywanie — pewne siebie, konkretne, bez „nie jestem pewien". Uczniowie się tego uczą. Jak się okazuje, modele również, bo oceniamy je w ten sam sposób. Autorzy przeanalizowali benchmarki, w których branża rzeczywiście rywalizuje, oraz rankingi, pod które dostraja modele. Prawie wszystkie przyznają odpowiedzi „nie wiem" dokładnie tyle samo punktów co błędnej odpowiedzi: zero. Przy takiej regule model, który zawsze zgaduje, pokona identyczny model uczciwie sygnalizujący niepewność. Całkiem dosłownie wtłaczamy je w halucynowanie za pomocą punktacji.

Dwa sposoby oceny odpowiedzi

co nagradza każda zasada punktacji

Zamknięte kryteria

tak punktuje dziś większość benchmarków

Poprawna	+1
Błędna	0
„Nie wiem"	0

Błędna odpowiedź i „Nie wiem" dają po 0, więc zgadywanie może tylko pomóc.

Jawne kryteria

punktacja podana w samym pytaniu

Poprawna	+1
Błędna	−1
„Nie wiem"	0

Błędne zgadywanie kosztuje teraz więcej niż uczciwe „Nie wiem".

Benchmarki mogą sprawić, że zgadywanie stanie się racjonalne. W typowym systemie punktacji (po lewej) zarówno błędna odpowiedź, jak i uczciwe „nie wiem" dają zero, więc zgadywanie może tylko pomóc. Gdy zasady podano w pytaniu i błędna odpowiedź wiąże się z karą (po prawej), wstrzymanie się od odpowiedzi przy niepewności może być lepszym wyborem. Zmienia to zachętę tworzoną przez test, ale samo w sobie nie

rozwiązuje problemu halucynacji.

Ten fragment warto zapamiętać, bo przeczy typowym nagłówkom. Halucynacje często przedstawia się jako nieuchronne, niemal mistyczne ograniczenie technologii. Praca kwestionuje oba określenia. Dolna granica błędów z pretreningu nie jest tajemnicą, lecz zwykłym błędem statystycznym, znanym uczeniu maszynowemu od dekad. Ich utrzymywanie się również nie jest nieuniknione: częściowo wynika z zachęty, którą sami stworzyliśmy i możemy zmienić. System, który po prostu odmawiałby odpowiedzi w razie niepewności, nie halucynowałby w ogóle. Wdrożone modele nie zachowują się tak, ponieważ nasze rankingi karzą odmowę.

Co zrobili autorzy

Praca składa się z trzech części. Pierwsza to argument matematyczny pokazujący, że pewna liczba halucynacji jest statystycznie wymuszona podczas pretreningu: zadanie „generuj wyłącznie poprawny tekst" jest co najmniej tak trudne jak binarna klasyfikacja „czy to zdanie jest prawdziwe?". Druga część — poparta przeglądem dziesięciu wpływowych benchmarków — dowodzi, że popularne miary dokładności nagradzają zgadywanie bardziej niż wstrzymanie się od odpowiedzi. Trzecia przedstawia proponowaną poprawkę i studium przypadku: oceny z **jawnymi kryteriami**, w których zasady punktacji umieszcza się w samym pytaniu, na przykład: „poprawna odpowiedź daje 1 punkt, błędna −1, więc zrezygnuj, jeśli twoja pewność jest mniejsza niż 50%". Dzięki temu model wie, kiedy uczciwość jest nagradzana. Autorzy sprawdzają to na czterech czołowych modelach — Gemini 3 Pro Google'a, GPT-5 OpenAI, Grok 4 firmy xAI i Claude Opus 4.5 firmy Anthropic — przy użyciu 4326 pytań faktograficznych z SimpleQA. Wyraźnie zaznaczają, że jest to ilustracyjne studium przypadku, a „nie kontrolowana ocena porównawcza modeli": użyto ustawień domyślnych, bez strojenia i bez normalizacji kosztów.

Co znaleźli

- **Pretrening wymusza pewien poziom błędów.** Częstość generowania przez model pewnie brzmiących fałszów ma dolną granicę równą w przybliżeniu dwukrotności błędu najlepszego klasyfikatora „czy to zdanie jest prawdziwe?" zbudowanego na jego podstawie. Dla faktów pozbawionych możliwego do nauczenia wzorca granicą jest co najmniej *odsetek singletonów* — udział faktów występujących w treningu dokładnie raz. Pewna liczba halucynacji jest nieunikniona nawet przy idealnie czystych danych.
- **Punktacja konkretnie nagradza zgadywanie.** Przy zwykłym podziale na odpowiedzi poprawne i błędne optymalną strategią jest nigdy się nie wstrzymywać, a przegląd autorów pokazuje, że zdecydowana większość popularnych benchmarków traktuje „nie wiem" po prostu jako błąd. Wyrazisty przykład dotyczy modeli samego OpenAI: w SimpleQA surowa dokładność nieznacznie *faworyzuje* o4-mini, który odpowiada niemal zawsze i myli się w ponad trzech czwartych przypadków, względem GPT-5-mini, który popełnia znacznie mniej błędów, ponieważ

w razie niepewności rezygnuje z odpowiedzi. Bardziej lekkomyślny model wygląda lepiej w rankingu.

- **Jawne kryteria odwracają zachętę w studium przypadku.** Autorzy testują prostą metodę ograniczania halucynacji: model odpowiada dwa razy i rezygnuje, jeśli odpowiedzi się różnią. Przy standardowej dokładności metoda zmniejsza liczbę błędów, ale także obniża wynik dokładności, więc miara zniechęca do jej stosowania. Przy jawnych kryteriach ta sama metoda wypada lepiej dla wszystkich czterech modeli w szerokim zakresie kar. GPT-5-mini, wcześniej karany przez surową dokładność za wstrzymywanie się od odpowiedzi, wyprzedza o4-mini, gdy zasady punktacji są jawne (n = 4326 pytań na model).

Co to prawdopodobnie oznacza

Ograniczanie halucynacji nie polega przede wszystkim na wymyślaniu kolejnych testów poświęconych wyłącznie halucynacjom. Trzeba zmienić sposób, w jaki główne benchmarki oceniają niepewność, aby przyznanie „nie wiem" nie było karane. Dopóki ranking się nie zmieni, ograniczanie halucynacji nadal będzie kosztować punkty dokładności, a więc pozostanie zniechęcane. Dlatego autorzy nazywają problem „społeczno-technicznym": potrzebna jest zarówno lepsza miara, jak i przekonanie wpływowych rankingów do jej przyjęcia.

Czego to *nie* dowodzi

- Praca **nie** pokazuje, że jawne kryteria naprawiają halucynacje w rzeczywistych zastosowaniach. Eksperyment wspierający tę tezę to małe, celowo **niekontrolowane** studium przypadku: cztery modele z ustawieniami domyślnymi, jedna wybrana metoda i jeden test pytań faktograficznych. Ma ono pokazać *odwrócenie zachęty*, a nie uszeregować modele czy dowieść ogólnej skuteczności.
- **Nie** twierdzi, że punktacja jest *jedyną* przyczyną. Błędy w danych treningowych, rzeczywiście trudne problemy i nietypowe polecenia pozostają odrębnymi źródłami.
- **Nie** wspiera popularnego twierdzenia, że halucynacje są nieuniknione. Autorzy argumentują odwrotnie: system odpowiadający tylko na sprawdzalne pytania, a w pozostałych przypadkach mówiący „nie wiem", nigdy by nie halucynował.
- **Nie** usuwa dolnej granicy błędów pretreningu — wyjaśnia ją i ogranicza liczbowo, przy czym granica dotyczy pewnie podawanych błędów faktograficznych, a nie wszystkich zachowań modelu.
- **Nie** pokazuje, że jawne kryteria są same w sobie *wystarczające*. Zmieniają to, co nagradza ocena, ale nie zastępują wyszukiwania informacji, używania narzędzi ani lepiej skalibrowanych modeli.

Jak mocne są dowody?

- Rdzeniem pracy jest **matematyka** — formalne dolne granice, a nie pomiary. Jako argument teoretyczny jest ona poprawna na własnych założeniach.
- Opiera się na **celowo uproszczonych modelach** problemu. Sami autorzy wskazują „fałszywą trychotomię" polegającą na traktowaniu każdej odpowiedzi jako poprawnej, błędnej albo „nie wiem", oraz wyidealizowany zbiór „arbitralnych faktów" użyty do uzyskania najczystszej granicy.
- Przegląd benchmarków obejmuje **małą, dobraną próbę** — dziesięć wpływowych ocen, a nie wyczerpujący audyt.
- Studium przypadku jest **rzeczywiste, ale ograniczone**: cztery czołowe modele, jedna metoda, tylko SimpleQA, ustawienia domyślne i wyraźne zastrzeżenie, że „nie jest to kontrolowana ocena". To dowód koncepcji dla argumentu o zachętach, a nie wynik rankingu.
- Warto nazwać perspektywę autorów: trzech z czterech pracuje lub pracowało w **OpenAI**, a artykuł przekonuje branżę do zmiany sposobu oceny modeli. To dobrze uzasadnione stanowisko zainteresowanej strony, nie neutralny zewnętrzny przegląd — należy je rozważyć, nie odrzucać. Na korzyść autorów przemawia to, że krytykują własne modele, o4-mini i GPT-5-mini, równie chętnie jak modele innych firm.

Dlaczego to ma znaczenie

Praca zmienia sposób myślenia o mocno nagłośnionym problemie. „Halucynację" przedstawia się zwykle jako upiorną wadę albo nieprzekraczalną ścianę; tutaj staje się ona czymś zwyczajnym i częściowo wywołanym przez nas samych — zrozumiałą granicą statystyczną, na którą nakłada się wybrany przez nas system zachęt. Szersza lekcja jest spokojniejsza i bardziej użyteczna: dalszy postęp w niezawodności może zależeć równie mocno od *tego, co mierzymy*, jak od tego, co budujemy.

Zwięzłe podsumowanie

Pewnie podawane fałszywe odpowiedzi modeli językowych mają dwa źródła. Pierwsze jest statystyczne: jeśli za faktem nie kryje się wzorzec, model wytrenowany do naśladowania języka czasami go pomyli, a dolną granicę błędów można oszacować na przykład na podstawie liczby faktów występujących w treningu tylko raz. Drugie źródło to zachęty: prawie każdy benchmark używany do tworzenia rankingów przyznaje „nie wiem" tyle samo punktów co błędnej odpowiedzi, więc zgadywanie zawsze wygrywa — nawet model mylący się w trzech czwartych przypadków może pokonać uczciwszy model, który rezygnuje z odpowiedzi. Propozycją autorów nie jest kolejny test halucynacji, lecz „jawne kryteria": podanie zasad punktacji w pytaniu. W studium przypadku obejmującym cztery czołowe modele odwróciło to zachętę, dzięki czemu metoda ograniczająca halucynacje została nagrodzona zamiast ukarana. Jest to recenzowana w *Nature* praca łącząca teorię, przegląd i mały, wyraźnie niekontrolowany eksperyment. Rozwiązanie jest obiecujące, ale jego skuteczności na dużą skalę jeszcze nie wykazano; autorzy dowodzą zaś, że halucynacje nie są ani tajemnicze, ani ściśle

nieuniknione.

Kontrola bez upiększeń

Co pokazuje praca: Matematyczną dolną granicę, przy której część halucynacji jest wymuszona podczas pretreningu — dla faktów bez wzorca co najmniej równą odsetkowi singletonów; przegląd wskazujący, że większość czołowych benchmarków nie przyznaje punktów za „nie wiem”; oraz studium czterech modeli, w którym podanie punktacji w poleceniu, czyli zastosowanie jawnych kryteriów, sprawia, że metoda ograniczająca halucynacje wygrywa, choć zwykła dokładność ją karała.

Co jest wiarygodne, ale nieudowodnione: Że wprowadzenie jawnych kryteriów do głównych benchmarków istotnie ograniczyłoby halucynacje we wdrożonych modelach. Wspierający eksperyment jest mały i wyraźnie niekontrolowany.

Czego praca nie pokazuje: Że halucynacje są nieuniknione — argumentuje przeciwnie; że punktacja jest jedyną przyczyną; że można całkowicie wyeliminować halucynacje; że studium przypadku stanowi ranking czterech modeli.

Główne ograniczenia: Celowo uproszczony podział na odpowiedzi poprawne, błędne i „nie wiem”, który autorzy nazywają „fałszywą trychotomią”; mały, dobrany przegląd dziesięciu benchmarków; niekontrolowane studium czterech modeli, jednej metody i jednego testu przy ustawieniach domyślnych; oraz fakt, że argument na temat oceny całej branży wyszedł głównie od badaczy związanych z OpenAI.

Jak duże zaufanie powinien mieć czytelnik niespecjalista? Wysokie, że halucynacje nie są ani tajemnicze, ani ściśle nieuniknione oraz że główne benchmarki obecnie nagradzają zgadywanie. Umiarkowane, że proponowana poprawka pomoże: istnieje rzeczywisty dowód koncepcji, ale nie wykazano jeszcze szerokiej skuteczności w dużej skali.

Źródła

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