The Clean Paper

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

Warum Sprachmodelle halluzinieren — und warum unsere Benotung es so hält

Die selbstbewussten Falschaussagen, die wir „Halluzinationen" nennen, sind kein mysteriöser Defekt: Ein Teil ist ein statistisches Nebenprodukt des Trainings, und sie halten sich, weil gängige Benchmarks einen selbstbewussten Rateversuch höher belohnen als ein ehrliches „ich weiß es nicht". Eine Fallstudie an vier Frontier-Modellen zeigt, dass das Nennen der Bewertungsregeln im Prompt („offene Rubriken") diesen Anreiz umkehrt.

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

Warum Modelle raten — und warum wir es ihnen beigebracht haben

Fragen Sie ein großes Sprachmodell nach dem Geburtstag eines Fremden, und es antwortet vielleicht „7. März" — mit der ruhigen Gewissheit von jemandem, der es von einer Karte abliest — und liegt falsch, dreimal hintereinander, mit drei verschiedenen Daten. Die Autoren geben genau diese Art Beispiel: führende Modelle, denen eine schlichte Faktenfrage gestellt wurde — der Geburtstag einer Person oder wofür ein obskures Akronym steht —, erfanden selbstbewusst jeweils eine andere Antwort, keine davon richtig. Das Wort der Branche dafür ist *Halluzination*, was nach einem Wahrnehmungsfehler klingt. Der erste Zug des Papers besteht darin, dem Phänomen das Mysterium zu nehmen.

Beginnen wir damit, wie ein Modell gebaut wird. In seiner ersten und größten Trainingsphase lernt es, im Kern, wie flüssige Sprache aussieht, indem es eine enorme Menge Text liest. Nun nehme man ein Faktum ohne dahinterliegendes Muster — den Geburtstag einer bestimmten Person. Wenn dieses Datum im Trainingstext einmal vorkam, oder nie, gibt es für einen Musterlerner nichts zu greifen: Die Antwort ist, aus Sicht des Modells, willkürlich. Die Autoren machen das präzise, indem sie eine alte Idee ausleihen (von Alan Turing, für ein anderes Problem entwickelt): Wenn eines von fünf Geburtsdaten in den Daten nur einmal auftaucht, ist zu erwarten, dass ein Modell mindestens eines von fünf falsch beantwortet — nicht weil es kaputt wäre, sondern weil dort nie etwas zu lernen war. (Nach derselben Logik verfehlen Modelle die Hauptstadt eines Landes fast nie: Die kommen ständig vor.) Sie argumentieren, mit einiger Sorgfalt, dass es selbst ein hartes Problem ist, eine wahre Aussage von einer plausiblen falschen zu unterscheiden — und dass es mindestens ebenso schwer ist, *nur* wahre Aussagen zu produzieren. Ein Boden an Fehlern ist eingebaut.

Die Idee darunter: wie man zählt, was man noch nicht gesehen hat

Das ruht auf einer wirklich klugen Idee — älter als Sprachmodelle, und es lohnt sich, ihr richtig zu begegnen.

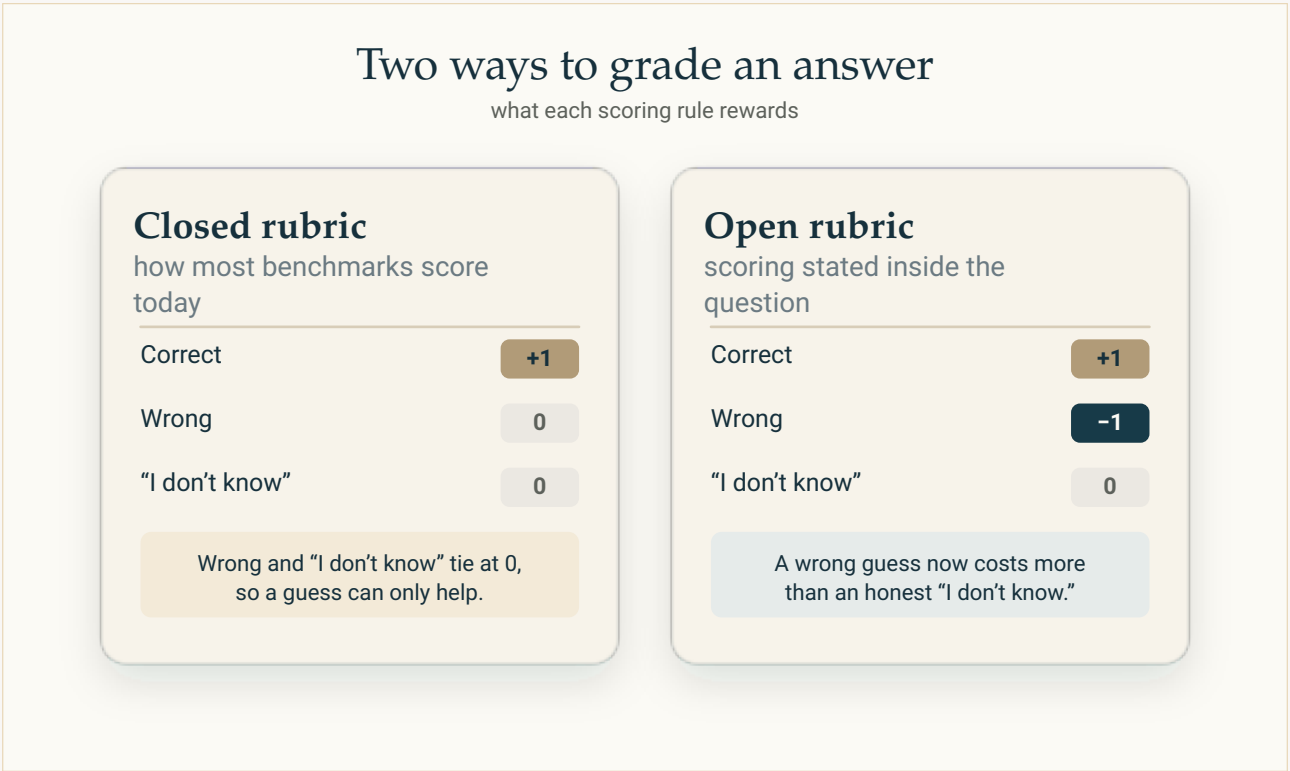
Man beginne mit einem Beutel bunter Kugeln. Sie wissen nicht, wie viele Farben er enthält. Sie ziehen 100, eine nach der anderen, und führen eine Strichliste: Rot 40, Blau 25, Grün 15, Gelb 5, Violett 3, Orange 2 — und dann **zehn verschiedene Farben, die je genau einmal auftauchen.**

Nun die Frage, vor der Turing tatsächlich stand, bei einem ganz anderen Problem: *Wie groß ist die Chance, dass die nächste Kugel eine Farbe hat, die Sie noch gar nicht gesehen haben?* Man kann nicht zählen, was man nie gezogen hat — aber man **kann** die Farben zählen, die man genau einmal gesehen hat, die „Singletons". Der Kniff, **Good-Turing-Schätzung** genannt, besteht darin, dass der Anteil der Züge, die Singletons sind, die Wahrscheinlichkeit schätzt, die noch in den ungesehenen Farben verborgen liegt. Zehn Ihrer hundert Züge waren Nur-einmal-Farben, also liegt die Chance, dass die nächste Kugel eine brandneue Farbe hat, bei etwa **10 / 100 = 10 %.**

Diese einmal gesehenen Farben sind keine *Fehler*. Sie sind eine Messung der eigenen Unwissenheit: Viele Farben, die nur einmal auftauchen, sind die Art der Stichprobe, einem mitzuteilen, dass die Welt mehr enthält, als man bislang gezogen hat.

> **Nun tausche man Farben gegen Geburtstage** und den Beutel gegen den Trainingstext des Modells. Angenommen, unter den Geburtstagen, die es sah, kommt **einer von fünf genau einmal vor.** Derselbe Kniff: Etwa ein Fünftel der Wahrscheinlichkeit liegt in Geburtstagen, die das Modell praktisch nie gesehen hat — und ein Geburtstag hat kein Muster, auf das man zurückfallen könnte (man kann sich den Geburtstag eines Menschen nicht *herleiten*). Ein einmal oder nie gesehenes Datum ist also eine Münze, die das Modell nicht gewichten kann, und es wird bei rund **einem von fünf** danebenliegen. Keine Klugheit der Welt behebt das: Da war nichts zu lernen.
> Das ist das ganze Argument in Miniatur: Die Singleton-Rate misst, wie viel von der Welt *aus diesen Daten* unlernbar ist, und daraus wird ein **Boden** unter den Fehlern. Deshalb verfehlt ein Modell auch fast nie eine Hauptstadt — *Paris* kommt ständig vor, seine Singleton-Rate liegt nahe null, es gibt also reichlich zu lernen.

Das erklärt, woher Halluzinationen kommen. Es erklärt nicht, warum sie *überleben* — warum Modelle nach all dem späteren Training, das sie hilfreich und ehrlich machen soll, immer noch bluffen, statt Zweifel einzugestehen. Hier ist die Analogie des Papers fast unangenehm treffend. Man stelle sich einen Studenten in einer Prüfung vor, der eine Antwort nicht weiß. Wenn ein leeres Feld null Punkte bringt und ein Rateversuch vielleicht einen, dann ist der notenmaximierende Zug: raten — selbstbewusst, konkret, niemals „ich bin nicht sicher". Studierende lernen das. Und Modelle, so zeigt sich, auch — weil wir sie genauso benoten. Die Autoren gingen die Benchmarks durch, auf denen das Feld tatsächlich konkurriert, die Ranglisten, für die Modelle getrimmt werden, und fanden: Fast alle geben „ich weiß es nicht" exakt dieselbe Punktzahl wie eine falsche Antwort — null. Unter dieser Regel schlägt ein Modell, das immer rät, ein ansonsten identisches Modell, das seine Unsicherheit ehrlich anzeigt. Wir benoten sie, ziemlich wörtlich, in das Verhalten hinein.



Benchmarks können Raten rational machen. Unter der Bewertung, die die meisten Benchmarks verwenden (links), bringen eine falsche Antwort und ein ehrliches „ich weiß es nicht" beide null Punkte — ein Rateversuch kann also nur helfen. Nennt man die Regeln in der Frage selbst, mit einem Abzug für Falsches (rechts), kann Enthaltung bei Unsicherheit zum besseren Zug werden. Das ändert, was der Test belohnt; es löst Halluzination nicht von allein.

Das ist der Teil, den man behalten sollte, denn er läuft der üblichen Schlagzeile zuwider. Halluzination wird oft als unvermeidliche, beinahe mystische Grenze der Technologie verkauft. Das Paper bestreitet beides. Der Pretraining-Boden ist kein Mysterium — er ist gewöhnlicher statistischer Fehler, wie ihn das maschinelle Lernen seit Jahrzehnten versteht. Und das Fortbestehen ist nicht unvermeidlich — es ist, zum Teil, ein Anreiz, den wir gebaut haben und ändern könnten. Ein System, das bei Unsicherheit schlicht die Antwort verweigerte, würde überhaupt nicht halluzinieren; dass ausgelieferte Modelle sich nicht so verhalten, liegt daran, dass unsere Anzeigetafeln die Verweigerung bestrafen.

Was die Autoren getan haben

Das Paper hat drei Teile. Erstens ein mathematisches Argument, dass ein Teil der Halluzination beim Pretraining statistisch erzwungen ist — gezeigt wird, dass „erzeuge nur gültigen Text" mindestens so schwer ist wie ein binäres Klassifikationsproblem „ist diese Aussage gültig¿'. Zweitens ein Argument — gestützt auf eine Durchsicht von zehn einflussreichen Benchmarks —, dass gängige Genauigkeitsmetriken Raten gegenüber Enthaltung belohnen. Drittens ein Lösungsvorschlag samt Fallstudie: **offene Rubriken** („open rubrics"), bei denen die Bewertung in der Frage selbst genannt wird (etwa: „eine richtige Antwort zählt 1, eine falsche −1, enthalte dich also, wenn du dir zu weniger als 50 % sicher bist"), sodass ein Modell erkennen kann, wann Ehrlichkeit belohnt wird. Sie erproben das an vier Frontier-Modellen — Googles Gemini 3 Pro, OpenAIs GPT-5, xAIs Grok 4 und Anthropics Claude Opus 4.5 — mit den 4326 Faktenfragen von SimpleQA. Sie sagen ausdrücklich, die Fallstudie sei illustrativ, „keine kontrollierte Evaluation über Modelle hinweg" (Standardeinstellungen, kein Tuning, keine Kostennormalisierung).

Was sie fanden

- **Pretraining erzwingt einen Teil der Fehler.** Die Rate, mit der ein Modell selbstbewusste Falschaussagen produziert, ist nach unten beschränkt durch (grob das Doppelte von) der Fehlerrate des besten daraus gebauten „Ist diese Aussage gültig?"-Klassifikators. Für Fakten ohne lernbares Muster liegt dieser Boden mindestens bei der *Singleton-Rate* — dem Anteil der Fakten, die genau einmal im Training vorkommen. Ein Teil der Halluzination ist selbst bei perfekt sauberen Daten unvermeidbar.
- **Die Benotung belohnt Raten — ganz konkret.** Unter gewöhnlicher Richtig/falsch-Bewertung ist Niemals-Enthalten die optimale Strategie, und die Durchsicht der Autoren findet, dass die

große Mehrheit der populären Benchmarks „ich weiß es nicht" schlicht als falsch wertet. Ein anschauliches Beispiel aus dem eigenen Haus: Auf dem SimpleQA-Test *begünstigt* die rohe Genauigkeit OpenAIs o4-mini leicht — das fast alles beantwortet und in mehr als drei Vierteln der Fälle falsch liegt — gegenüber GPT-5-mini, das weit weniger Fehler macht, weil es sich bei Unsicherheit enthält. Das rücksichtslosere Modell sieht auf der Anzeigetafel besser aus.

- **Offene Rubriken drehen den Anreiz um (in ihrer Fallstudie).** Getestet wird eine einfache Halluzinations-Milderung (das Modell zweimal antworten lassen und sich enthalten, wenn die beiden Antworten nicht übereinstimmen). Unter Standard-Genauigkeit senkt die Milderung die Fehler, *senkt aber auch die Genauigkeit* — die Metrik rät also von ihr ab. Unter offenen Rubriken schneidet dieselbe Milderung bei allen vier Modellen über eine Spanne von Strafwerten hinweg besser ab; und GPT-5-mini — das die rohe Genauigkeit für sein Enthalten bestraft hatte — liegt vor o4-mini, sobald die Bewertung offen genannt wird (n = 4326 Fragen pro Modell).

Was das wahrscheinlich bedeutet

Halluzination zu verringern ist überwiegend keine Frage weiterer halluzinationsspezifischer Tests. Es ist eine Frage dessen, wie die gängigen Benchmarks Unsicherheit bewerten — sodass das Eingeständnis „ich weiß es nicht" nicht länger bestraft wird. Solange sich die Anzeigetafel nicht ändert, wird das Verringern von Halluzination Modelle weiter Genauigkeitspunkte kosten und damit weiter entmutigt werden — weshalb die Autoren das Problem als „sozio-technisch" rahmen: teils bessere Metrik, teils die einflussreichen Ranglisten dazu bringen, sie zu übernehmen.

Was das *nicht* beweist

- Es zeigt **nicht**, dass offene Rubriken Halluzination in freier Wildbahn beheben. Das stützende Experiment ist eine kleine, absichtlich **unkontrollierte** Fallstudie — vier Modelle mit Standardeinstellungen, eine gewählte Milderung, ein Fakten-QA-Test —, gedacht, den *Anreizwechsel* zu demonstrieren, nicht Modelle zu ranken oder allgemeine Wirksamkeit zu belegen.
- Es behauptet **nicht**, die Benotung sei die *einzige* Ursache. Fehler in den Trainingsdaten, wirklich schwere Probleme und unvertraute Prompts bleiben eigenständige Quellen.
- Es stützt **nicht** die verbreitete Linie, Halluzinationen seien unvermeidlich. Die Autoren argumentieren das Gegenteil: Ein System, das nur überprüfbare Fragen beantwortete und ansonsten „ich weiß es nicht" sagte, würde nie halluzinieren.
- Es lässt den Pretraining-Boden **nicht** verschwinden — es erklärt und beschränkt ihn, und die Schranke betrifft selbstbewusste Faktenfehler, nicht jedes Modellverhalten.
- Es zeigt **nicht**, dass offene Rubriken für sich genommen *ausreichen*. Sie ändern, was eine Evaluation belohnt; sie sind kein Ersatz für Retrieval, Werkzeugnutzung oder besser kalibrierte Modelle.

Wie belastbar sind die Belege?

- Der Kern ist **Mathematik** — formale untere Schranken, keine Messungen. Als theoretisches Argument ist er in seinen eigenen Begriffen stichhaltig.
- Er ruht auf **bewusst vereinfachten Modellen** des Problems; die Autoren selbst benennen die „falsche Trichotomie", jede Antwort als richtig, falsch oder „ich weiß es nicht" zu behandeln, und das idealisierte Setting „willkürlicher Fakten" für die sauberste Schranke.
- Die Benchmark-Durchsicht ist eine **kleine, kuratierte Stichprobe** — zehn einflussreiche Evaluationen, kein erschöpfendes Audit.
- Die Fallstudie ist **real, aber begrenzt**: vier Frontier-Modelle, eine einzige Milderung, nur SimpleQA, Standardeinstellungen, ausdrücklich „keine kontrollierte Evaluation". Sie ist ein Konzeptnachweis für das Anreiz-Argument, kein Benchmark-Ergebnis.
- Der Blickwinkel gehört benannt: Drei der vier Autoren sind oder waren bei **OpenAI** beschäftigt, und das Paper argumentiert, das Feld solle ändern, wie es Modelle bewertet. Das ist eine gut begründete Position einer interessierten Partei, keine neutrale Außensicht — zu gewichten, nicht abzutun. (Ehrenhalber: Das Paper richtet die Kritik ebenso bereitwillig auf die eigenen Modelle, o4-mini und GPT-5-mini, wie auf andere.)

Warum das wichtig ist

Es rahmt ein stark gehyptes Problem neu. „Halluzination" wird gern entweder als unheimlicher Defekt oder als unverrückbare Wand verkauft; dieses Paper macht sie gewöhnlich und teilweise selbstverschuldet — ein statistischer Boden, den wir tatsächlich verstehen können, auf einem Anreiz, den wir gewählt haben. Die weitere Lehre ist leiser und nützlicher: Weiterer Fortschritt bei der Verlässlichkeit könnte ebenso sehr davon abhängen, *was wir messen*, wie davon, was wir bauen.

Kurz zusammengefasst

Selbstbewusste falsche Antworten von Sprachmodellen kommen aus zwei Quellen. Die erste ist statistisch: Wenn ein Faktum kein lernbares Muster hat, wird ein Modell, das Sprache imitieren soll, es manchmal verfehlen, und dieser Boden lässt sich abschätzen (etwa daran, wie viele Fakten nur einmal im Training vorkommen). Die zweite sind Anreize: Fast jeder Benchmark, nach dem Modelle gerankt werden, wertet „ich weiß es nicht" wie eine falsche Antwort, sodass Raten immer gewinnt — bis zu dem Punkt, dass ein Modell, das in drei Vierteln der Fälle falsch liegt, ein ehrlicheres überholen kann, das sich enthält. Der Vorschlag der Autoren ist kein weiterer Halluzinationstest, sondern „offene Rubriken": die Bewertung in der Frage selbst nennen. In einer Fallstudie an vier Frontier-Modellen dreht das den Anreiz um, sodass eine halluzinationssenkende Methode belohnt statt bestraft wird. Es ist ein Theorie-plus-Durchsicht-Paper mit einem kleinen, ausdrücklich unkontrollierten Experiment, begutachtet in *Nature* erschienen; der Fix ist vielversprechend, aber noch nicht im großen Maßstab erwiesen, und Halluzinationen sind, so das Argument, weder mysteriös noch

strikt unvermeidlich.

Nüchtern geprüft

Was das Paper zeigt: Eine mathematische untere Schranke, unter der ein Teil der Halluzination beim Pretraining erzwungen ist (mindestens die „Singleton-Rate" bei musterlosen Fakten); eine Durchsicht, die findet, dass die meisten führenden Benchmarks „ich weiß es nicht" keinerlei Punkte geben; und eine Fallstudie an vier Modellen, in der das Nennen der Bewertung im Prompt („offene Rubriken") eine halluzinationssenkende Methode gewinnen lässt, wo rohe Genauigkeit sie bestraft hatte.

Was plausibel, aber nicht bewiesen ist: Dass offene Rubriken, in gängige Benchmarks übernommen, Halluzination in ausgelieferten Modellen spürbar verringern würden. Das stützende Experiment ist klein und ausdrücklich unkontrolliert.

Was es nicht zeigt: Dass Halluzinationen unvermeidlich sind (es argumentiert das Gegenteil); dass die Benotung die einzige Ursache ist; dass Halluzination sich restlos beseitigen lässt; dass die Fallstudie die vier Modelle gegeneinander rankt.

Wesentliche Einschränkungen: Ein bewusst vereinfachtes Richtig/falsch/„ich weiß es nicht"-Modell (die Autoren nennen es eine „falsche Trichotomie"); eine kleine kuratierte Benchmark-Durchsicht (zehn Evaluationen); eine unkontrollierte Fallstudie (vier Modelle, eine Milderung, ein Test, Standardeinstellungen); und ein OpenAI-geführtes Argument darüber, wie das Feld Modelle bewerten sollte.

Wie viel Vertrauen sollte eine allgemeine Leserin haben? Hohes darin, dass Halluzination weder mysteriös noch strikt unvermeidlich ist und dass gängige Benchmarks derzeit Raten belohnen. Mäßiges darin, dass der vorgeschlagene Fix hilft — er hat jetzt einen echten Konzeptnachweis, aber noch keinen Beleg, dass er breit und im großen Maßstab funktioniert.

Quellen

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