The Clean Paper

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

Por qué los modelos de lenguaje alucinan, y por qué nuestra forma de evaluarlos mantiene el problema

Las falsedades dichas con seguridad que llamamos "alucinaciones" no son un fallo misterioso: algunas son un subproducto estadístico del entrenamiento, y persisten porque los benchmarks dominantes premian una conjetura segura más que un honesto "no lo sé". Un estudio de caso con cuatro modelos de frontera muestra que declarar las reglas de puntuación en el prompt ("rúbricas abiertas") invierte ese incentivo.

Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso · 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

Por qué los modelos adivinan, y por qué les enseñamos a hacerlo

Pídele a un gran modelo de lenguaje la fecha de nacimiento de un desconocido y puede contestar "7 de marzo" con la confianza firme de alguien que la lee de una tarjeta — y estar equivocado, tres veces seguidas, con tres fechas distintas. Los autores dan exactamente este tipo de ejemplo: modelos líderes ante una pregunta factual sencilla — la fecha de nacimiento de una persona, o qué significa un acrónimo oscuro — inventan cada uno una respuesta distinta con seguridad, y ninguna es correcta. La palabra de la industria para esto es *alucinación*, que lo hace sonar como un fallo de percepción. El primer movimiento del artículo es quitarle el misterio.

Empieza por cómo se construye un modelo. En su primera y mayor etapa de entrenamiento aprende, en la práctica, cómo suena el lenguaje fluido leyendo una cantidad enorme de texto. Ahora toma un hecho sin patrón detrás: la fecha de nacimiento de una persona concreta. Si esa fecha apareció en el texto de entrenamiento una vez, o nunca, no hay nada a lo que un aprendiz de patrones pueda agarrarse: desde el punto de vista del modelo, la respuesta es arbitraria. Los autores lo hacen preciso tomando prestada una vieja idea (de Alan Turing, para otro problema): si una de cada cinco fechas de nacimiento aparece solo una vez en los datos, deberíamos esperar que un modelo falle al menos una de cada cinco — no porque esté roto, sino porque nunca hubo nada que aprender. (Por la misma lógica, los modelos casi nunca fallan la capital de un país: esos datos aparecen constantemente.) Argumentan, con cuidado, que distinguir una afirmación verdadera de una falsa pero plausible es ya un problema difícil, y que producir *solo* afirmaciones verdaderas es al menos igual de difícil. Hay un suelo de error incorporado.

La idea de fondo: cómo contar lo que aún no has visto

Esto descansa en una idea realmente inteligente, más antigua que los modelos de lenguaje, y vale la pena verla bien.

Empieza con una bolsa de bolas de colores. No sabes cuántos colores contiene. Sacas 100, una por una, y las cuentas: rojo 40, azul 25, verde 15, amarillo 5, morado 3, naranja 2 — y luego **diez colores distintos que aparecen exactamente una vez cada uno.**

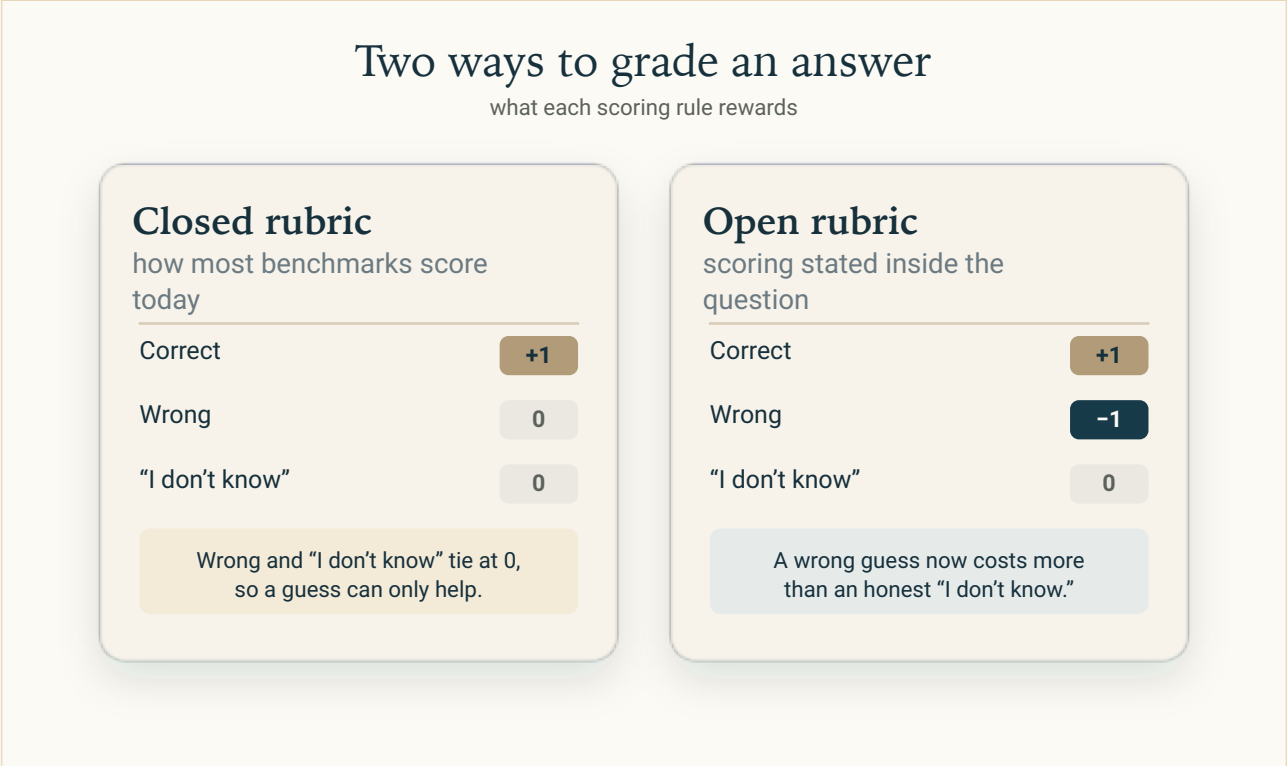
Ahora la pregunta que Turing enfrentaba realmente, en un problema muy distinto: *¿cuál es la probabilidad de que la próxima bola sea de un color que no has visto todavía?* No puedes contar lo que nunca has sacado, pero **sí** puedes contar los colores que has visto exactamente una vez, los "singletons". El truco, llamado **estimación de Good-Turing**, es que la parte de tus extracciones que son singletons estima la probabilidad que sigue escondida en los colores que no has visto. Diez de tus cien extracciones fueron colores vistos una sola vez, así que la probabilidad de que la próxima bola sea de un color completamente nuevo es de alrededor de **10 / 100 = 10%.**

Esos colores vistos una vez no son *errores*. Son una medición de tu propia ignorancia: muchos colores que aparecen una sola vez son la forma en que la muestra te dice que el mundo contiene más cosas que simplemente aún no has sacado.

Ahora cambia colores por cumpleaños, y la bolsa por el texto de entrenamiento del modelo. Supón que, entre los cumpleaños que vio, **uno de cada cinco aparece exactamente una vez.** El mismo truco: alrededor de un quinto de la probabilidad vive en cumpleaños que el modelo, en

> la práctica, nunca ha visto — y un cumpleaños no tiene un patrón al que recurrir (no puedes *calcular* el cumpleaños de alguien). Así que una fecha vista una vez, o nunca, es una moneda que el modelo no puede ponderar, y se equivocará en aproximadamente **una de cada cinco**. Ninguna astucia arregla esto: no había nada que aprender.
> Ese es todo el argumento en miniatura: la tasa de singletons mide cuánto del mundo es imposible de aprender *a partir de estos datos*, y eso se convierte en un **suelo** bajo los errores. También explica por qué un modelo casi nunca falla una capital: *París* aparece constantemente, su tasa de singletons está cerca de cero, así que hay mucho que aprender.

Eso explica de dónde vienen las alucinaciones. No explica por qué *sobreviven*: por qué los modelos, después de todo el entrenamiento posterior destinado a hacerlos útiles y honestos, siguen faroleando en vez de admitir la duda. Aquí la analogía del artículo es casi incómodamente acertada. Imagina a un estudiante en un examen que no sabe una respuesta. Si dejarla en blanco puntúa cero y adivinar podría dar un punto, el movimiento que maximiza la nota es adivinar: con confianza, de forma específica, nunca "no estoy seguro". Los estudiantes aprenden esto. Los modelos también, al parecer, porque los puntuamos de la misma manera. Los autores revisaron los benchmarks en los que el campo compite realmente, los rankings que los modelos se ajustan para escalar, y encontraron que casi todos dan a "no lo sé" exactamente la misma puntuación que a una respuesta equivocada: cero. Con esa regla, un modelo que siempre adivina superará a un modelo por lo demás idéntico que marca honestamente su incertidumbre. En un sentido bastante literal, los estamos puntuando hacia eso.



Los benchmarks pueden hacer racional la conjetura. Con la puntuación que usan la mayoría de los benchmarks (izquierda), una respuesta incorrecta y un honesto "no lo sé" valen ambos cero, así que adivinar solo puede ayudar. Si se declaran las reglas en la pregunta, con una penalización por equivocarse (derecha), abstenerse cuando hay duda puede convertirse en la mejor opción. Cambia lo que la prueba recompensa; por sí solo, no

resuelve la alucinación.

Esta es la parte que conviene retener, porque va contra el titular habitual. La alucinación suele venderse como un límite inevitable, casi místico, de la tecnología. El artículo discute ambas cosas. El suelo del preentrenamiento no es un misterio: es error estadístico ordinario, del tipo que el aprendizaje automático entiende desde hace décadas. Y la persistencia no es inevitable: es, en parte, un incentivo que construimos y que podríamos cambiar. Un sistema que simplemente se negara a responder cuando no está seguro no alucinaría en absoluto; la razón por la que los modelos desplegados no se comportan así es que nuestros marcadores castigan la negativa.

Qué hicieron los autores

El artículo tiene tres partes. Primero, un argumento matemático de que cierta alucinación es estadísticamente forzada durante el preentrenamiento, mostrando que "generar solo texto válido" es al menos tan difícil como un problema binario de clasificación: "¿esta afirmación es válida?". Segundo, un argumento — apoyado por una revisión de diez benchmarks influyentes — de que las métricas corrientes de tipo exactitud recompensan adivinar por encima de abstenerse. Tercero, una propuesta de arreglo y un estudio de caso para probarlo: evaluaciones de **rúbrica abierta**, donde la puntuación se declara dentro de la propia pregunta (por ejemplo, "una respuesta correcta puntúa 1, una equivocada -1, así que abstente si estás menos del 50% seguro"), para que un modelo pueda saber cuándo se recompensa la honestidad. Lo prueban en cuatro modelos de frontera — Gemini 3 Pro de Google, GPT-5 de OpenAI, Grok 4 de xAI y Claude Opus 4.5 de Anthropic — usando las 4.326 preguntas factuales de SimpleQA. Son explícitos en que el estudio de caso es ilustrativo, "no una evaluación controlada entre modelos" (configuraciones por defecto, sin ajuste, sin normalización de costes).

Qué encontraron

- **El preentrenamiento fuerza cierto error.** La tasa con la que un modelo emite falsedades confiadas está acotada por abajo por (aproximadamente el doble de) la tasa de error del mejor clasificador "¿esta afirmación es válida?" construido a partir de él. Para hechos sin patrón aprendible, ese suelo es al menos la *tasa de singletons*: la fracción de hechos que aparecen exactamente una vez en el entrenamiento. Alguna alucinación es inevitable incluso con datos perfectamente limpios.
- **La puntuación recompensa adivinar, concretamente.** Bajo una puntuación ordinaria correcto/incorrecto, no abstenerse nunca es la estrategia óptima, y la revisión de los autores encuentra que la gran mayoría de benchmarks populares puntúan "no lo sé" simplemente como incorrecto. Un ejemplo vívido de su propio lado: en la prueba SimpleQA, la exactitud bruta *favorece* ligeramente a o4-mini de OpenAI — que responde casi todo y se equivoca más de tres cuartas partes de las veces — frente a GPT-5-mini, que comete muchos menos errores porque se abstiene cuando no está seguro. El modelo más temerario se ve mejor en el marcador.
- **Las rúbricas abiertas invierten el incentivo (en su estudio de caso).** Prueban una mitigación

sencilla de la alucinación (hacer que el modelo responda dos veces y se abstenga si las dos respuestas no coinciden). Bajo exactitud estándar, la mitigación reduce los errores *pero también reduce la exactitud*, así que la métrica desincentiva adoptarla. Bajo rúbricas abiertas, la misma mitigación sale ganando para los cuatro modelos a través de un rango de penalizaciones; y GPT-5-mini — al que la exactitud bruta había penalizado por abstenerse cuando no estaba seguro — queda por delante de o4-mini una vez que la puntuación se declara abiertamente (n = 4.326 preguntas por modelo).

Qué significa probablemente

Reducir la alucinación no es sobre todo cuestión de inventar más pruebas específicas de alucinación. Es cuestión de cambiar cómo los benchmarks dominantes puntúan la incertidumbre, para que admitir "no lo sé" deje de ser castigado. Mientras el marcador no cambie, reducir la alucinación seguirá costando puntos de exactitud a los modelos y, por tanto, seguirá desincentivándose; por eso los autores enmarcan el problema como "sociotécnico": en parte una métrica mejor, en parte lograr que los rankings influyentes la adopten.

Qué no prueba

- No muestra que las rúbricas abiertas arreglen la alucinación en el mundo real. El experimento de apoyo es un estudio de caso pequeño, deliberadamente **no controlado** — cuatro modelos con configuraciones por defecto, una mitigación elegida, una prueba de preguntas factuales — destinado a demostrar el *cambio de incentivo*, no a clasificar modelos ni a probar eficacia general.
- No afirma que la puntuación sea la *única* causa. Los errores en los datos de entrenamiento, los problemas realmente difíciles y los prompts desconocidos siguen siendo fuentes separadas.
- No apoya la idea popular de que las alucinaciones son inevitables. Los autores sostienen lo contrario: un sistema que respondiera solo a preguntas verificables y dijera "no lo sé" en los demás casos nunca alucinaría.
- No hace desaparecer el suelo del preentrenamiento: lo explica y lo acota, y la cota se refiere a errores factuales confiados, no a todo el comportamiento del modelo.
- No muestra que las rúbricas abiertas sean *suficientes* por sí solas. Cambian lo que una evaluación recompensa; no sustituyen la recuperación de información, el uso de herramientas ni modelos mejor calibrados.

¿Qué tan sólida es la evidencia?

- El núcleo es **matemático**: cotas inferiores formales, no mediciones. Como argumento teórico, es sólido bajo sus propios supuestos.
- Descansa en modelos del problema **deliberadamente simplificados**; los propios autores señalan

la “falsa tricotomía” de tratar toda respuesta como correcta, incorrecta o “no lo sé”, y el escenario idealizado de “hechos arbitrarios” usado para la cota más limpia.
- La revisión de benchmarks es una **muestra pequeña y seleccionada**: diez evaluaciones influyentes, no una auditoría exhaustiva.
- El estudio de caso es **real pero limitado**: cuatro modelos de frontera, una sola mitigación, solo SimpleQA, configuraciones por defecto, explícitamente “no una evaluación controlada”. Es una prueba de concepto para el argumento de incentivos, no un resultado de benchmark.
- Vale la pena nombrar el punto de vista: tres de los cuatro autores trabajan o trabajaron en **OpenAI**, y el artículo argumenta que el campo debería cambiar cómo evalúa los modelos. Es una posición bien argumentada de una parte interesada, no una revisión externa neutral: hay que sopesarla, no descartarla. (Para su crédito, el artículo dirige la crítica a sus propios modelos, o4-mini y GPT-5-mini, con la misma facilidad que a otros.)

Por qué importa

Reencuadra un problema cargado de hype. La “alucinación” suele venderse como un defecto inquietante o como un muro inmóvil; este artículo la vuelve ordinaria y en parte autoinfligida: un suelo estadístico que podemos entender de verdad, encima de un incentivo que elegimos. La lección más amplia es más discreta y más útil: el progreso adicional en fiabilidad puede depender tanto de *lo que medimos* como de lo que construimos.

Resumen limpio

Las respuestas falsas y confiadas de los modelos de lenguaje vienen de dos lugares. El primero es estadístico: cuando un hecho no tiene patrón que aprender, un modelo entrenado para imitar lenguaje a veces lo fallará, y ese suelo puede estimarse (por ejemplo, a partir de cuántos hechos aparecen solo una vez en el entrenamiento). El segundo son los incentivos: casi todos los benchmarks en los que se clasifican los modelos puntúan “no lo sé” igual que una respuesta equivocada, así que adivinar siempre gana, hasta el punto de que un modelo equivocado tres cuartas partes del tiempo puede superar a otro más honesto que se abstiene. La propuesta de los autores no es otra prueba de alucinación sino “rúbricas abiertas”: declarar la puntuación dentro de la pregunta. En un estudio de caso con cuatro modelos de frontera, eso invierte el incentivo, de modo que se recompensa un método que reduce la alucinación en vez de penalizarlo. Es un artículo de teoría más revisión con un experimento pequeño y explícitamente no controlado, revisado por pares en *Nature*; el arreglo es prometedor pero aún no se ha demostrado a gran escala, y las alucinaciones se presentan como ni misteriosas ni estrictamente inevitables.

Sin rodeos

Qué muestra el artículo: Una cota inferior matemática según la cual cierta alucinación es forzada durante el preentrenamiento (al menos la "tasa de singletons" para hechos sin patrón); una revisión que encuentra que la mayoría de benchmarks importantes no dan crédito a "no lo sé"; y un estudio de caso de cuatro modelos en el que declarar la puntuación en el prompt ("rúbricas abiertas") hace ganar a un método reductor de alucinaciones, donde la exactitud simple lo había penalizado.

Qué es plausible pero no está probado: Que las rúbricas abiertas, incorporadas a los benchmarks dominantes, reducirían de forma significativa la alucinación en modelos desplegados. El experimento de apoyo es pequeño y explícitamente no controlado.

Qué no muestra: Que las alucinaciones sean inevitables (sostiene lo contrario); que la puntuación sea la única causa; que la alucinación pueda eliminarse por completo; que el estudio de caso clasifique los cuatro modelos entre sí.

Principales limitaciones: Un modelo deliberadamente simplificado correcto/incorrecto/"no lo sé" (los autores lo llaman una "falsa tricotomía"); una revisión pequeña y seleccionada de benchmarks (diez evaluaciones); un estudio de caso no controlado (cuatro modelos, una mitigación, una prueba, configuraciones por defecto); y un argumento liderado por OpenAI sobre cómo el campo debería evaluar los modelos.

¿Cuánta confianza debería tener un lector general? Alta en que la alucinación no es ni misteriosa ni estrictamente inevitable, y en que los benchmarks dominantes hoy recompensan adivinar. Moderada en que la solución propuesta ayuda: ya tiene una verdadera prueba de concepto, pero todavía no una demostración de que funcione de forma amplia y a escala.

Sources

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