The Clean Paper

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

Perché i modelli linguistici allucinano, e perché il modo in cui li valutiamo mantiene il problema

Le falsità sicure di sé che chiamiamo "allucinazioni" non sono un malfunzionamento misterioso: alcune sono un sottoprodotto statistico dell'addestramento, e persistono perché i benchmark più usati premiano una risposta tirata a indovinare con sicurezza più di un onesto "non lo so". Un caso di studio su quattro modelli frontier mostra che dichiarare nel prompt le regole di valutazione ("rubriche aperte") inverte questo incentivo.

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

Perché i modelli tirano a indovinare, e perché glielo abbiamo insegnato

Chiedi a un modello linguistico di grandi dimensioni il compleanno di una persona sconosciuta e potrebbe rispondere “7 marzo” con la sicurezza tranquilla di chi sta leggendo da una scheda — e sbagliarsi, tre volte di fila, con tre date diverse. Gli autori danno proprio esempi di questo tipo: modelli di punta a cui viene posta una domanda fattuale semplice — il compleanno di una persona, o il significato di un acronimo poco noto — inventano ogni volta una risposta diversa con sicurezza, e nessuna è corretta. L’industria chiama questo fenomeno *allucinazione*, una parola che lo fa sembrare un difetto di percezione. La prima mossa del lavoro è togliergli il mistero.

Partiamo da come viene costruito un modello. Nella prima e più grande fase di addestramento impara, in pratica, che aspetto ha il linguaggio fluente leggendo enormi quantità di testo. Ora prendiamo un fatto che non ha uno schema dietro di sé — il compleanno di una persona specifica. Se quella data è comparsa una sola volta nel testo di addestramento, o non è comparsa affatto, non c’è nulla a cui un sistema che apprende schemi possa aggrapparsi: dal punto di vista del modello, la risposta è arbitraria. Gli autori rendono precisa l’idea prendendo in prestito un vecchio argomento (di Alan Turing, nato per un problema diverso): se un compleanno su cinque compare una sola volta nei dati, ci si dovrebbe aspettare che un modello sbagli almeno uno su cinque di quei compleanni — non perché sia rotto, ma perché non c’era nulla da imparare. (Per la stessa logica, i modelli sbagliano quasi mai la capitale di un paese: quelle compaiono continuamente.) Gli autori argomentano, con cautela, che distinguere un’affermazione vera da una falsa ma plausibile è già un problema difficile, e che produrre *solo* affermazioni vere è almeno altrettanto difficile. Esiste un livello minimo di errore.

L’idea sotto: come contare ciò che non hai ancora visto

Qui sotto c’è un’idea davvero elegante — più vecchia dei modelli linguistici, e vale la pena incontrarla con calma.

Immagina un sacchetto di palline colorate. Non sai quanti colori contenga. Ne estrai 100, una alla volta, e le conti: rosse 40, blu 25, verdi 15, gialle 5, viola 3, arancioni 2 — e poi **dieci colori diversi che compaiono una sola volta ciascuno.**

Ora la domanda che Turing si trovò davvero davanti, in un problema molto diverso: *qual è la probabilità che la prossima pallina sia di un colore che non hai ancora visto?* Non puoi contare ciò che non hai mai estratto — ma puoi contare i colori che hai visto esattamente una volta, i “singleton”. Il trucco, chiamato **stima di Good–Turing**, è che la quota delle tue estrazioni composta da singleton stima la probabilità ancora nascosta nei colori che non hai visto. Dieci delle cento estrazioni erano colori comparsi una sola volta, quindi la probabilità che la prossima pallina sia di un colore nuovo è circa **10 / 100 = 10%.**

Quei colori visti una volta sola non sono *errori*. Sono una misura della tua ignoranza: molti colori che compaiono una sola volta sono il modo in cui il campione ti dice che il mondo contiene altro che semplicemente non hai ancora estratto.

Ora sostituisci i colori con i compleanni, e il sacchetto con il testo di addestramento del modello. Supponiamo che, tra i compleanni che ha visto, **uno su cinque compaia esattamente una volta.** Stesso trucco: circa un quinto della probabilità vive in compleanni che il modello,

> di fatto, non ha mai visto — e un compleanno non ha uno schema su cui ripiegare (non puoi *dedurre* il compleanno di qualcuno). Quindi una data vista una volta sola, o mai vista, è una moneta che il modello non può pesare, e sbaglierà su circa **uno su cinque** di quei casi. Nessuna astuzia risolve questo punto: non c'era nulla da imparare.
> Questo è l'argomento in miniatura: il tasso di singleton misura quanta parte del mondo è non apprendibile *da questi dati*, e questo diventa un **pavimento** sotto gli errori. È anche il motivo per cui un modello quasi non sbaglia mai una capitale: *Parigi* compare continuamente, il suo tasso di singleton è vicino a zero, quindi c'è molto da imparare.

Questo spiega da dove vengono le allucinazioni. Non spiega perché *sopravvivano* — perché i modelli, dopo tutto l'addestramento successivo che dovrebbe renderli utili e onesti, continuino a bluffare invece di ammettere il dubbio. Qui l'analogia del lavoro è quasi scomoda per quanto è adatta. Immagina uno studente in un esame che non sa una risposta. Se lasciare in bianco vale zero e tirare a indovinare può valere uno, la mossa che massimizza il voto è indovinare — con sicurezza, nello specifico, mai "non ne sono sicuro". Gli studenti lo imparano. A quanto pare lo imparano anche i modelli — perché li valutiamo nello stesso modo. Gli autori hanno passato in rassegna i benchmark su cui il settore compete davvero, le classifiche che i modelli vengono ottimizzati per scalare, e hanno trovato che quasi tutti danno a "non lo so" lo stesso punteggio di una risposta sbagliata: zero. Con questa regola, un modello che prova sempre a indovinare batterà un modello altrimenti identico che segnala onestamente la propria incertezza. In senso abbastanza letterale, li stiamo valutando dentro questo comportamento.

Due modi di valutare una risposta

che cosa premia ogni regola di punteggio

Rubrica chiusa

come valuta oggi la maggior parte dei benchmark

Corretta	+1
Sbagliata	0
«Non lo so»	0

Sbagliata e «non lo so» valgono 0: indovinare può solo aiutare.

Rubrica aperta

punteggio dichiarato nella domanda

Corretta	+1
Sbagliata	−1
«Non lo so»	0

Una risposta sbagliata ora costa più di un onesto «non lo so».

I benchmark possono rendere razionale tirare a indovinare. Con il punteggio usato dalla maggior parte dei benchmark (a sinistra), una risposta sbagliata e un onesto "non lo so" valgono entrambi zero — quindi

indovinare può solo aiutare. Se invece le regole sono dichiarate nella domanda, con una penalità per l'errore (a destra), astenersi quando non si è sicuri può diventare la scelta migliore. Cambia ciò che il test premia; da solo non risolve le allucinazioni.

Original diagram — The Clean Paper CC BY 4.0

Questa è la parte da tenere stretta, perché va contro il titolo facile. L'allucinazione viene spesso presentata come un limite inevitabile, quasi mistico, della tecnologia. Il lavoro contesta entrambe le cose. Il pavimento dell'addestramento iniziale non è un mistero: è normale errore statistico, del tipo che il machine learning conosce da decenni. E la persistenza non è inevitabile: è, in parte, un incentivo che abbiamo costruito e che potremmo cambiare. Un sistema che rifiutasse semplicemente di rispondere quando non è sicuro non allucinerebbe affatto; il motivo per cui i modelli dispiegati non si comportano così è che le nostre classifiche puniscono il rifiuto.

Che cosa hanno fatto gli autori

Il lavoro ha tre parti. Primo, un argomento matematico secondo cui una certa quantità di allucinazione è statisticamente forzata durante il preaddestramento, mostrando che "generare solo testo valido" è almeno difficile quanto un problema binario di classificazione "questa affermazione è valida?". Secondo, un argomento — sostenuto da una rassegna di dieci benchmark influenti — secondo cui le metriche più diffuse basate sull'accuratezza premiano l'ipotesi azzardata rispetto all'astensione. Terzo, una proposta di correzione e un caso studio che la testa: valutazioni a **rubrica aperta**, in cui il punteggio viene dichiarato dentro la domanda stessa (per esempio, "una risposta corretta vale 1, una sbagliata −1, quindi astieniti se sei sicuro meno del 50%"), così che un modello possa capire quando l'onestà viene premiata. Lo provano su quattro modelli di frontiera — Gemini 3 Pro di Google, GPT-5 di OpenAI, Grok 4 di xAI e Claude Opus 4.5 di Anthropic — usando le 4.326 domande fattuali di SimpleQA. Gli autori sono espliciti: il caso studio è illustrativo, "non una valutazione controllata tra modelli" (impostazioni predefinite, nessuna ottimizzazione, nessuna normalizzazione dei costi).

Che cosa hanno trovato

- **Il preaddestramento impone una certa quantità di errore.** Il tasso con cui un modello emette falsità sicure ha un limite inferiore pari, grossomodo, al doppio del tasso di errore del miglior classificatore "questa affermazione è valida?" costruito a partire da esso. Per fatti senza uno schema apprendibile, quel pavimento è almeno il *tasso di singleton* — la frazione di fatti che compaiono esattamente una volta nell'addestramento. Una certa allucinazione è inevitabile anche con dati perfettamente puliti.
- **La valutazione premia concretamente il tirare a indovinare.** Con il normale punteggio corretto/sbagliato, non astenersi mai è la strategia ottimale, e la rassegna degli autori mostra che la grande maggioranza dei benchmark popolari valuta "non lo so" semplicemente come sbagliato. Un esempio vivido dai loro dati: su SimpleQA, l'accuratezza grezza *favorisce leggermente* o4-mini di OpenAI — che risponde quasi sempre ed è sbagliato più di tre quarti delle volte — rispetto a GPT-5-

mini, che fa molti meno errori perché si astiene quando non è sicuro. Il modello più imprudente appare migliore in classifica.
- **Le rubriche aperte ribaltano l'incentivo (nel loro caso studio).** Testano una mitigazione semplice dell'allucinazione (far rispondere il modello due volte e farlo astenere se le due risposte non coincidono). Con l'accuratezza standard, la mitigazione riduce gli errori *ma riduce anche l'accuratezza* — quindi la metrica scoraggia l'adozione. Con le rubriche aperte, la stessa mitigazione risulta vantaggiosa per tutti e quattro i modelli in un intervallo di penalità; e GPT-5-mini — che l'accuratezza grezza aveva penalizzato perché si asteneva quando era incerto — supera o4-mini una volta che il punteggio viene dichiarato apertamente (n = 4.326 domande per modello).

Che cosa probabilmente significa

Ridurre le allucinazioni non è soprattutto una questione di inventare altri test specifici per l'allucinazione. È una questione di cambiare il modo in cui i benchmark principali valutano l'incertezza, in modo che ammettere "non lo so" non venga più punito. Finché la classifica non cambia, ridurre le allucinazioni continuerà a costare punti di accuratezza ai modelli e quindi continuerà a essere scoraggiato — ed è per questo che gli autori descrivono il problema come "socio-tecnico": in parte metrica migliore, in parte convincere le classifiche influenti ad adottarla.

Che cosa questo non dimostra

- Non dimostra che le rubriche aperte risolvano le allucinazioni nel mondo reale. L'esperimento a supporto è un caso studio piccolo e deliberatamente **non controllato** — quattro modelli con impostazioni predefinite, una mitigazione scelta, un test di domande fattuali — pensato per mostrare il *ribaltamento dell'incentivo*, non per classificare i modelli o provare un'efficacia generale.
- Non sostiene che la valutazione sia l'*unica* causa. Errori nei dati di addestramento, problemi realmente difficili e prompt non familiari restano fonti separate.
- Non supporta la frase popolare secondo cui le allucinazioni sono inevitabili. Gli autori argomentano il contrario: un sistema che rispondesse solo a domande verificabili e altrimenti dicesse "non lo so" non allucinerebbe mai.
- Non fa sparire il limite minimo imposto dal preaddestramento: lo spiega e lo delimita, e quel limite riguarda errori fattuali sicuri, non tutto il comportamento dei modelli.
- Non mostra che le rubriche aperte siano *sufficienti* da sole. Cambiano ciò che una valutazione premia; non sostituiscono retrieval, uso di strumenti o modelli meglio calibrati.

Quanto sono solide le prove?

- Il nucleo è **matematico** — limiti inferiori formali, non misurazioni. Come argomento teorico regge nei propri termini.
- Si appoggia a **modelli volutamente semplificati** del problema; gli stessi autori segnalano la "falsa tricotomia" di trattare ogni risposta come corretta, scorretta o "non lo so", e l'ambientazione idealizzata dei "fatti arbitrari" usata per il limite più pulito.
- La rassegna dei benchmark è un **campione piccolo e selezionato** — dieci valutazioni influenti, non un audit esaustivo.
- Il caso studio è **reale ma limitato**: quattro modelli di frontiera, una sola mitigazione, solo SimpleQA, impostazioni predefinite, esplicitamente "non una valutazione controllata". È una prova di concetto per l'argomento sugli incentivi, non un risultato da classifica.
- Vale la pena nominare il punto di vista: tre dei quattro autori sono o sono stati dipendenti di **OpenAI**, e il lavoro sostiene che il settore dovrebbe cambiare il modo in cui valuta i modelli. È una posizione ben argomentata da una parte interessata, non una revisione esterna neutrale — da pesare, non da liquidare. (A suo merito, il lavoro rivolge la critica ai propri modelli, o4-mini e GPT-5-mini, tanto quanto agli altri.)

Perché conta

Riformula un problema molto gonfiato dal marketing. "Allucinazione" tende a essere venduta o come un difetto inquietante o come un muro invalicabile; questo lavoro la rende ordinaria e in parte autoinflitta — un pavimento statistico che possiamo capire davvero, sopra un incentivo che abbiamo scelto. La lezione più ampia è più quieta e più utile: il progresso sull'affidabilità potrebbe dipendere tanto da *che cosa misuriamo* quanto da che cosa costruiamo.

In breve

Le risposte false e sicure dei modelli linguistici vengono da due fonti. La prima è statistica: quando un fatto non ha pattern da imparare, un modello addestrato a imitare il linguaggio a volte sbaglierà, e quel pavimento può essere stimato (per esempio, da quanti fatti compaiono una sola volta nell'addestramento). La seconda sono gli incentivi: quasi ogni benchmark su cui i modelli vengono classificati dà a "non lo so" lo stesso punteggio di una risposta sbagliata, quindi indovinare conviene sempre — al punto che un modello sbagliato tre quarti delle volte può superare un modello più onesto che si astiene. La proposta degli autori non è un altro test per le allucinazioni ma una "rubrica aperta": dichiarare il punteggio dentro la domanda. In un caso studio su quattro modelli di frontiera, questo ribalta l'incentivo e premia un metodo che riduce le allucinazioni invece di penalizzarlo. È un lavoro teorico più una rassegna più un piccolo esperimento esplicitamente non controllato, sottoposto a revisione tra pari su *Nature*; la correzione è promettente ma non è ancora dimostrato che funzioni in modo ampio e su scala, e le allucinazioni vengono descritte come né

misteriose né strettamente inevitabili.

Valutazione critica

Che cosa mostra il lavoro: un limite inferiore matematico secondo cui una certa quantità di allucinazione è forzata durante il preaddestramento (almeno il "tasso di singleton" per fatti senza pattern); una rassegna che trova che la maggior parte dei benchmark principali non dà credito a "non lo so"; e un caso studio su quattro modelli in cui dichiarare il punteggio nel prompt ("rubriche aperte") fa vincere un metodo che riduce le allucinazioni, mentre la semplice accuratezza lo penalizzava.

Che cosa è plausibile ma non dimostrato: che le rubriche aperte, integrate nei benchmark principali, ridurrebbero in modo significativo le allucinazioni nei modelli dispiegati. L'esperimento a supporto è piccolo ed esplicitamente non controllato.

Che cosa non mostra: che le allucinazioni siano inevitabili (sostiene il contrario); che la valutazione sia l'unica causa; che l'allucinazione possa essere eliminata del tutto; che il caso studio classifichi i quattro modelli tra loro.

Limiti principali: un modello volutamente semplificato corretto/sbagliato/"non lo so" (gli autori lo chiamano una "falsa tricotomia"); una piccola rassegna selezionata di benchmark (dieci valutazioni); un caso studio non controllato (quattro modelli, una mitigazione, un test, impostazioni predefinite); e un argomento guidato da OpenAI su come il settore dovrebbe valutare i modelli.

Quanta fiducia dovrebbe avere un lettore generale? Alta sul fatto che l'allucinazione non sia né misteriosa né strettamente inevitabile, e che i benchmark principali oggi premino il tirare a indovinare. Moderata sul fatto che la correzione proposta aiuti: ora ha una vera prova di concetto, ma non ancora una dimostrazione che funzioni in modo ampio e su scala.

Fonti

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