The CLEAN PAPER

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

Language models hallucinate क्योंकरते हैं — और हमारीgrading उन्हें ऐसाकरते रहने के लिए क्यों प्रेत्सहति करतीहै

जनि आत्मवश्विासीझूठे उत्तरोंकोहम "hallucinations" कहते हैं, वे कोई रहस्यमय glitch नहीहैं। कुछ training कीstatistical by-product हैं; और वे इसलिए बने रहते हैं क्योंकि mainstream benchmarks ईमानदार "मुझे नहीं पता" कीतुलनामें confident guess कोअधकि reward करते हैं। चार frontier models पर case study दखितीहै कि prompt में scoring rules साफ़ लखि देना("open rubrics") इस incentive कोउलट सकताहै।

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

मॉडल अनुमान क्यों लगाते हैं — और हमने उन्हें ऐसा करना क्यों सिखाया

किसी large language model से किसी अनजान व्यक्ति की जन्मतिथि पूछिए। वह “7 मार्च” ऐसे आत्मविश्वास से बता सकता है जैसे कार्ड से पढ़ रहा हो— और लगातार तीन बार तीन अलग तारीखें देकर तीनों बार गलत हो सकता है। लेखक ठीक ऐसा ही उदाहरण देते हैं: प्रमुख मॉडलों से किसी व्यक्ति की जन्मतिथि या किसी दुर्लभ acronym का पूरा रूप पूछा गया, और हर मॉडल ने आत्मविश्वास से अलग उत्तर गढ़ दिया, कोई सही नहीं था। उद्योग इसे *hallucination* कहता है, जिससे यह किसी धारणा-संबंधी खराबी जैसा सुनाई देता है। शोध-पत्र का पहला कदम इस रहस्य को साधारण बनाता है।

शुरुआत इस बात से करें कि मॉडल बनता कैसे है। प्रशिक्षण के पहले और सबसे बड़े चरण में वह बहुत विशाल पाठ-संग्रह पढ़कर, मोटे तौर पर, भाषा के पैटर्न सीखता है। अब ऐसी जानकारी लें जिसके पीछे कोई पैटर्न नहीं है — किसी खास व्यक्ति की जन्मतिथि। यदि वह तारीख प्रशिक्षण-पाठ में एक बार आई या बिल्कुल नहीं आई, तो pattern learner के पास पकड़ने के लिए कुछ नहीं है: मॉडल की दृष्टि से उत्तर मनमाना है। लेखक इसे एक पुराने विचार — अलग समस्या के लिए Alan Turing द्वारा इस्तेमाल विचार — से गणितीय रूप देते हैं। यदि पाँच में से एक जन्मतिथि डेटा में केवल एक बार दिखाई देती है, तो मॉडल से कम से कम पाँच में से एक ऐसी जन्मतिथि गलत होने की अपेक्षा करनी चाहिए — इसलिए नहीं कि मॉडल टूटा है, बल्कि इसलिए कि सीखने के लिए पर्याप्त सूचना की भी थी ही नहीं। इसी तर्क से देश की राजधानी जैसी बातें मॉडल शायद ही गलत करता है, क्योंकि वे पाठ में बार-बार आती हैं। लेखक सावधानी से तर्क देते हैं कि सत्य कथन को विश्वसनीय लेकिन असत्य कथन से अलग पहचानना स्वयं कठिन समस्या है, और *केवल* सत्य कथन उत्पन्न करना कम से कम उतना ही कठिन है। इसलिए त्रुटि की एक न्यूनतम सीमा प्रशिक्षण में ही मौजूद रहती है।

मूल विचार: जो आपने अभी तक देखा ही नहीं, उसे कैसे गिनें?

यह तर्क एक सचमुच चतुर विचार पर आधारित है — भाषा मॉडलों से कहीं पुराना, और अलग से समझने लायक।

रंगीन गेंदों से भरे थैले की कल्पना करें। आप को नहीं पता उसमें कितने रंग हैं। आप एक-एक करके 100 गेंदें निकालते हैं और गिनते हैं: लाल 40, नीली 25, हरी 15, पीली 5, बैंगनी 3, नारंगी 2 — और फिर **दस अलग रंग ऐसे हैं जो केवल एक-एक बार मिले।**

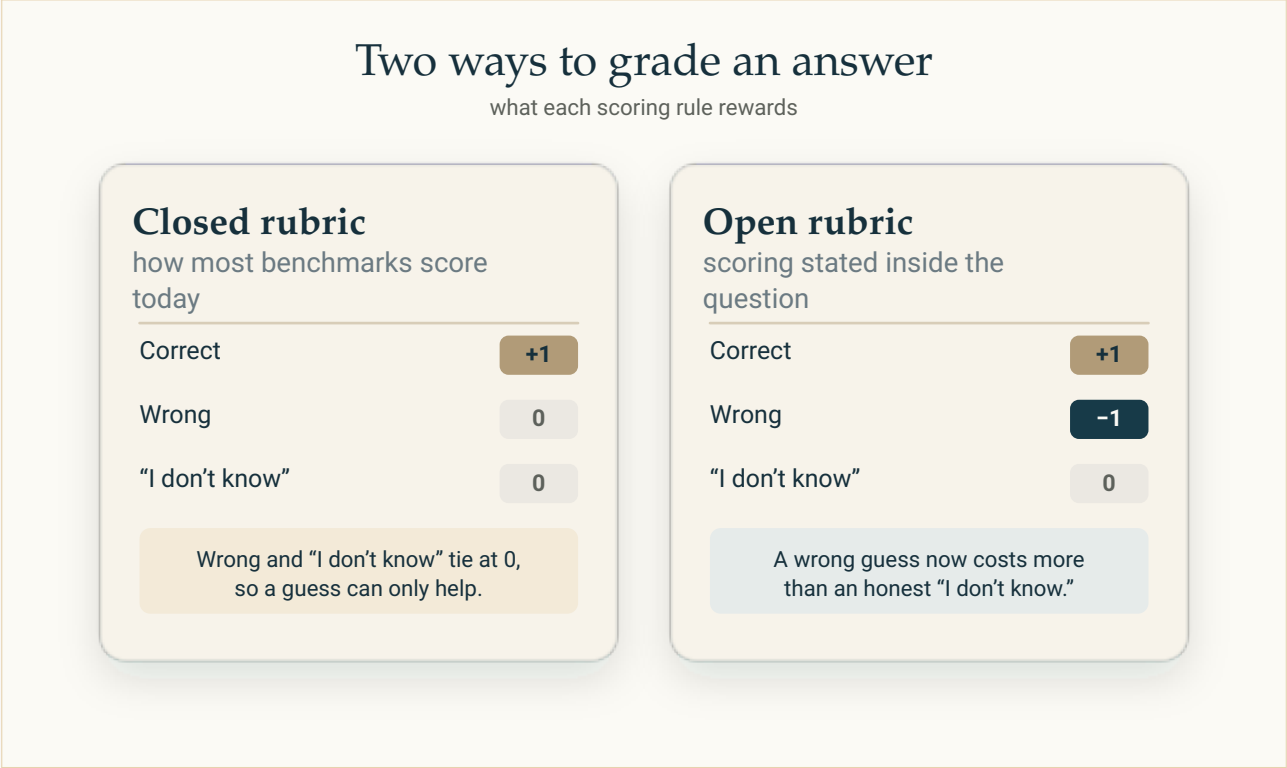
अब वह सवाल जिसे Turing ने एक अलग समस्या में पूछा था: *अगली गेंद का रंग ऐसा होने की संभावना कितनी है जो आपने अभी तक एक बार भी नहीं देखा?* जिसे कभी निकाला ही नहीं, उसे सीधे गिन नहीं सकते। लेकिन जो रंग **सिर्फ एक बार** मिले हैं — singletons — उन्हें गिन सकते हैं। **Good–Turing estimation** नाम की तरकीब कहती है कि नमूने में singletons का अनुपात लगभग उस संभावना का अनुमान देता है जो अभी तक न दिखे रंगों में छिपी है। 100 में 10 निकली गेंदें ऐसे रंगों की थीं जो केवल एक बार मिले, इसलिए अगली गेंद का बिल्कुल नया रंग होने की संभावना लगभग **10 / 100 = 10%** है।

एक बार मिले रंग *गलतियाँ* नहीं हैं। वे आपकी अपनी अज्ञानता का माप हैं: यदि बहुत-से रंग केवल एक बार दिखे, तो नमूना आपको बता रहा है कि दुनिया में ऐसे और रंग भी हैं जिन्हें आपने अभी तक निकाला ही नहीं।

अब रंगों की जगह जन्मतिथियाँ रखिए, और थैले की जगह मॉडल का प्रशिक्षण-पाठ। मान लें कि देखी गई जन्मतिथियों में **पाँच में से एक केवल एक बार दिखाई देती है।** वही तरकीब कहती है कि लगभग पाँचवाँ हिस्सा ऐसी जन्मतिथियों की संभावना में है जिन्हें मॉडल ने प्रभावी रूप से कभी ठीक से नहीं देखा। जन्मतिथि में अनुमान लगाने लायक कोई सामान्य पैटर्न भी नहीं होता। इसलिए एक बार या कभी न देखी गई तारीख पर मॉडल के पास सही उत्तर तक पहुँचने का भरोसेमंद आधार नहीं है, और लगभग **पाँच में से एक** पर गलती अपेक्षित है। केवल अधिक चतुर तर्क से इसे ठीक नहीं किया जा सकता; वहाँ सीखने के लिए पर्याप्त सूचना थी ही नहीं।

यही पूरे तर्क का छोटा रूप है: singleton rate बताती है कि इस डेटा से दुनिया का कितना हिस्सा सीखना संभव नहीं है, और वही त्रुटियों के नीचे एक **न्यूनतम सीमा** बन जाता है। इसी वजह से राजधानी जैसे तथ्य बहुत कम चूकते हैं — *Paris* जैसे नाम लगातार आते हैं, उनकी singleton rate लगभग शून्य है और सीखने के लिए बहुत कुछ मौजूद है।

यह बताता है कि hallucination पैदा कहाँ से होती है। लेकिन यह नहीं बताता कि बड़े के प्रशिक्षण के बावजूद वह *बची क्यों रहती* है — मॉडल संदेह मानने के बजाय आत्मविश्वास से अनुमान क्यों लगाता है। यहाँ शोधपत्र की उपमा असहज रूप से सटीक है। परीक्षा में ऐसे छात्र की कल्पना कीजिए जिसे उत्तर नहीं पता। यदि खाली छोड़ने पर शून्य अंक और अनुमान लगाने पर सही होने की थोड़ी भी संभावना है, तो अंक अधिकतम करने की रणनीति है: अनुमान लगाओ — आत्मविश्वास से, स्पष्ट उत्तर दो, कभी "मुझे नहीं पता" मत कहो। छात्र यह सीखते हैं। मॉडल भी सीखते हैं, क्योंकि हम उन्हें लगभग इसी तरह अंक देते हैं। लेखकों ने उन benchmarks और leaderboards को देखा जिन पर क्षेत्र प्रतिस्पर्धा करता है और जिनके अनुसार मॉडल optimize किए जाते हैं। लगभग सभी में "मुझे नहीं पता" को गलत उत्तर के समान ही अंक मिलता है: शून्य। ऐसे नियम में हमेशा अनुमान लगाने वाला मॉडल उस बिल्कुल समान मॉडल से बेहतर स्कोर करेगा जो अनिश्चित होने पर ईमानदारी से रुक जाता है। काफ़ी शब्दकि अर्थ में, हमारी scoring ही उन्हें अनुमान लगाने के लिए प्रेरित करती है।



Benchmarks अनुमान लगाना तर्कसंगत बनाते हैं। समान्य scoring में (बाएँ) गलत उत्तर और ईमानदार "मुझे नहीं पता" दोनों को शून्य मिलता है, इसलिए अनुमान लगाने से केवल फ़ायदा हो सकता है। यदि प्रश्न में scoring साफ़ लिखी जाए और गलत उत्तर पर दंड हो (दाएँ), तो अनिश्चित होने पर जवाब न देना बेहतर रणनीति बन सकती है। इससे परीक्षण का प्रोत्साहन बदलता है; अपने-आप hallucination की समस्या हल नहीं होती।

Original diagram — The Clean Paper CC BY 4.0

यही हिस्सा याद रखने लायक है, क्योंकि यह आम सुर्खी के विपरीत जाता है। Hallucination को अक्सर तकनीक की अनिवार्य, लगभग रहस्यमय सीमा की तरह पेश किया जाता है। शोधपत्र दोनों बातों को चुनौती देता है। Pretraining की न्यूनतम त्रुटि कोई रहस्य नहीं— यह समान्य सांख्यिकीय त्रुटि है, जिस तरह की चीज़ machine learning दशकों से समझती है। और बड़े में उसका बने रहना भी पूरी तरह अनिवार्य नहीं— उसका एक हिस्सा हमारी बनाई हुई incentive structure से आता है, जिसे बदला जा सकता है। यदि कोई प्रणाली अनिश्चित होने पर केवल उत्तर देने से मना कर दे, तो वह उस स्थिति में hallucinate नहीं करेगी। तैनात मॉडल ऐसा लगातार नहीं करते क्योंकि हमारे scoreboards अक्सर उस ईमानदार refusal को दंडित करते हैं।

लेखकों ने क्या किया

शोधपत्र के तीन हिस्से हैं। पहला गणितीय तर्क है कि pretraining के दौरान कुछ hallucination सांख्यिकीय रूप से अनिवार्य है: "केवल वैध पाठ उत्पन्न करना" कम से कम binary "क्या यह कथन वैध है?" classification जितना कठिन समस्या है। दूसरा तर्क, दस प्रभावशाली benchmarks के सर्वेक्षण से समर्थित, यह है कि मुख्यधारा के accuracy metrics जवाब न देने की तुलना में अनुमान लगाने को पुरस्कृत करते हैं। तीसरा एक प्रस्तावित सुधार और उसका छोटा case study है: **open-rubric evaluations**, जिनमें scoring नियम प्रश्न के भीतर ही साफ़ लिखे जाते हैं — जैसे "सही उत्तर +1, गलत −1; यदि 50% से कम भरोसा हो तो जवाब न दें" — ताकि मॉडल जान सकें कि ईमानदार अनिश्चितता को इनाम मिलेगा। इसे उन्होंने चार अग्रणी मॉडलों — Google Gemini 3 Pro, OpenAI GPT-5, xAI Grok 4 और Anthropic Claude Opus 4.5 — पर SimpleQA के 4,326 factual questions के साथ आज़माया। वे स्पष्ट लिखते हैं कि यह केवल उदाहरणात्मक case study है, "मॉडलों के बीच नियंत्रित मूल्यांकन नहीं" — default settings, कोई tuning नहीं और लगात का normalization नहीं।

उन्हें क्या मिला

- **Pretraining कुछ त्रुटि अनिवार्य बनाता है।** मॉडल द्वारा आत्मविश्वास से गलत तथ्य उत्पन्न करने की दर की एक निचली सीमा है, जो मोटे तौर पर उससे बने सर्वोत्तम "क्या यह कथन वैध है?" classifier की त्रुटि दर से जुड़ी है। जिन तथ्यों में सीखने लायक पैटर्न नहीं, वहाँ यह सीमा कम से कम *singleton rate* — प्रशिक्षण में केवल एक बार दिखाई देने वाले तथ्यों का अनुपात — जितनी होती है। पूरी तरह साफ़ डेटा होने पर भी कुछ factual hallucination से बचना संभव नहीं होगा।
- **Scoring अनुमान लगाने को सचमुच पुरस्कृत करती है।** सामान्य सही/गलत अंकन में कभी जवाब न छोड़ना सर्वोत्तम रणनीति बन जाती है, और लेखकों के सर्वेक्षण में लोकप्रिय benchmarks की बड़ी बहुमत "मुझे नहीं पता" को बस गलत मानती है। उनका स्पष्ट उदाहरण SimpleQA से है: raw accuracy OpenAI के o4-mini को थोड़ा बेहतर दिखाती है — जो लगभग हर सवाल का जवाब देता है और तीन-चौथाई से अधिक बार गलत होता है — GPT-5-mini की तुलना में, जो अनिश्चित होने पर अधिक बार रुकता है और इसलिए बहुत कम गलतियाँ करता है। अधिक जोखिम लेने वाला मॉडल scoreboard पर बेहतर दिखाई देता है।
- **Open rubrics incentive को पलट सकती हैं — कम से कम इस case study में।** लेखक सरल hallucination-mitigation आज़माते हैं: मॉडल से दो बार उत्तर लें और दो उत्तर अलग हों तो जवाब न दें। सामान्य accuracy में इससे गलतियाँ कम होती हैं, लेकिन accuracy score भी घटता है — यानी metric इस उपाय को अपनाने से हतोत्साहित करती है। Open rubrics में penalties की अलग-अलग मात्रा पर यही उपाय चारों मॉडलों के लिए बेहतर निकलता है; और GPT-5-mini, जिसे raw accuracy ने अधिक abstain करने के लिए दंडित किया था, scoring साफ़ बताए जाने पर o4-mini से आगे निकलता है। हर मॉडल के लिए n = 4,326 सवाल थे।

इसका शायद क्या मतलब है

Hallucination घटाने का बड़ा हिस्सा शायद और अधिक hallucination-specific tests बनाने में नहीं, बल्कि मुख्यधारा के benchmarks में अनिश्चितता को अंक देने का तरीका बदलने में है, ताकि "मुझे नहीं पता" कहना दंडित न हो। जब तक scoreboard नहीं बदलता, hallucination कम करने से मॉडल को accuracy points खोने पड़ सकते हैं और इसलिए ऐसी रणनीतियाँ अपनाने का incentive कमज़ोर रहेगा। इसी वजह से लेखक समस्या को "socio-technical" कहते हैं: कुछ हिस्सा बेहतर metric का है, और कुछ यह मनवाने का कि प्रभावशाली leaderboards उसे अपनाएँ।

यह क्या साबित नहीं करता

- यह **नहीं** दिखाता कि open rubrics वास्तविक दुनिया की hallucination समस्या हल कर देती हैं। सहायक प्रयोग छोटा और जानबूझकर **uncontrolled** case study है — चार मॉडल, default settings, एक चुना mitigation, एक factual-QA test — जिसका उद्देश्य *incentive बदलना* दिखाना है, मॉडलों की ranking या व्यापक प्रभावशीलता साबित करना नहीं।
- यह **नहीं** कहता कि scoring ही *एकमात्र* कारण है। Training data की त्रुटियाँ, सचमुच कठिन समस्याएँ और अपरिचित prompts अलग स्रोत बने रहते हैं।
- यह लोकप्रिय दावे का समर्थन **नहीं** करता कि hallucinations अनिवार्य हैं। लेखक उल्टा तर्क देते हैं: ऐसी प्रणाली जो केवल जाँचे जा सकने वाले सवालों का उत्तर दे और बाक़ी पर "मुझे नहीं पता" कहे, hallucinate नहीं करेगी।
- यह pretraining की न्यूनतम त्रुटि **गायब नहीं** करता। शोधपत्र उसे समझाता और सीमा देता है; वह सीमा आत्मविश्वास से दिए गए factual errors के बारे में है, मॉडल के हर व्यवहार के बारे में नहीं।
- यह **नहीं** दिखाता कि open rubrics अकेले पर्याप्त हैं। वे evaluation में मिलने वाले reward को बदलती हैं; retrieval, tool use या बेहतर calibrated models की जगह नहीं लेतीं।

सिद्धांत कितना मजबूत है?

- मुख्य आधार **गणित** है — औपचारिक lower bounds, न कि केवल मापे गए correlations। सैद्धांतिक तर्क अपनी धारणाओं के भीतर मजबूत है।
- यह समस्या के **जानबूझकर सरल किए गए मॉडल** पर टिका है। लेखक स्वयं हर उत्तर को सही, गलत या "मुझे नहीं पता" में बाँटने को "false trichotomy" कहते हैं, और सबसे साफ़ सीमा के लिए "arbitrary facts" वाली आदर्शीकृत परिस्थिति इस्तेमाल करते हैं।
- Benchmark survey **छोटा और चुनाहुआ नमूना** है — दस प्रभावशाली evaluations, पूरे क्षेत्र का exhaustive audit नहीं।
- Case study **वास्तविक लेकिन सीमित** है: चार frontier models, एक mitigation, केवल SimpleQA, default settings और स्पष्ट घोषणा कि यह "controlled evaluation" नहीं है। यह incentive तर्क का proof of concept है, benchmark result नहीं।
- दृष्टिकोण का स्रोत बताना भी उचित है: चार में से तीन लेखक **OpenAI** में काम करते हैं या कर चुके हैं, और शोधपत्र क्षेत्र को model evaluation बदलने की सलाह देता है। यह हितधारक पक्ष से आया सुविचारित तर्क है, निष्पक्ष बाहरी समीक्षा नहीं — इसे तौलना चाहिए, केवल इसी कारण खारिज नहीं। श्रेय की बात है कि शोधपत्र अपनी आलोचना o4-mini और GPT-5-mini जैसे OpenAI मॉडलों पर भी उतनी ही आसानी से लागू करता है।

यह क्यों मायने रखता है

यह बहुत hype वाली समस्या को नए ढंग से रखता है। "Hallucination" को अक्सर या तो रहस्यमय खराबी या अचल दीवार की तरह पेश किया जाता है; यह शोधपत्र इसे साधारण और आंशिक रूप से हमारी अपनी बनाई समस्या बनाता है — एक सांख्यिकीय न्यूनतम सीमा जिसे समझा जा सकता है, और उसके ऊपर एक incentive जिसे हमने चुना है। बड़ा सबक शांत लेकिन अधिक उपयोगी है: reliability में आगे की प्रगति उतनी ही *इस बात पर निर्भर हो सकती है कि हम क्या मापते हैं* जितनी इस पर कि हम क्या बनाते हैं।

संक्षेप में

भाषामॉडलोंके आत्मविश्वास से दिए गए गलत तथ्य दोस्रोतोंसे आते हैं। पहलासांख्यिकीय है: जिस तथ्य के पीछे सीखने लायक कोई पैटर्न नहीं, भाषाकीनकल सीखने वालामॉडल कभीकभीउसे गलत बताएगा; इस न्यूनतम त्रुटि काअनुमान लगायाजासकताहै, उदाहरण के लिए प्रशिक्षण में केवल एक बार दिखाई देने वाले तथ्योंके अनुपात से। दूसराincentive है: लगभग हर ऐसाbenchmark जिस पर मॉडलोंकीranking होतीहै, "मुझे नहींपता" कोगलत उत्तर के बराबर अंक देताहै, इसलिए अनुमान लगानाहमेशाफायदे कासौदाहै — इतनाकि तीन-चौथाई से अधिक बार गलत होने वालामॉडल भीअधिक ईमानदारीसे abstain करने वाले मॉडल से बेहतर raw score पासकताहै। लेखकोंकाप्रस्ताव कोई नयाhallucination test नहीं, बल्कि "open rubrics" है: scoring नियम प्रश्न में हीसाफ़ लिखें। चार frontier models पर छोटे case study में इससे incentive बदल गयाऔर hallucination घटाने वालातरीकादंडित होने के बजाय पुरस्कृत हुआ। यह theory + survey वालाशोधपत्र है, जिसके साथ छोटाऔर स्पष्ट रूप से uncontrolled experiment है। प्रस्ताव आशाजनक है, लेकिन अभीबड़े पैमाने पर सिद्ध नहीं, और मुख्य तर्क यह है कि hallucination न रहस्यमय है, न पूरीतरह अनिवार्य।

बिनाबढ़ाचढ़ाकर जाँच

शोधपत्र क्यादिखाताहै: गणितीय lower bound जिसके तहत pretraining में कुछ hallucination अनिवार्य है — pattern-less facts के लिए कम से कम singleton rate के स्तर पर; survey जिसमें अधिकांश प्रमुख benchmarks "मुझे नहींपता" कोकोई लाभ नहींदेते; और चार मॉडलोंवालाcase study जिसमें prompt के भीतर scoring साफ़ लिखने — open rubrics — पर hallucination घटाने वालातरीकाजीतताहै, जबकि साधारण accuracy उसे दंडित करतीथी।

क्यासंभव है, पर साबित नहीं कि open rubrics कोमुख्यधाराके benchmarks में अपनाने से तैनात मॉडलोंकीhallucination अर्थपूर्ण रूप से घटेगी। सहायक प्रयोग छोटाऔर स्पष्ट रूप से uncontrolled है।

यह क्यानहींदिखाता कि hallucinations पूरीतरह अनिवार्य हैं — शोधपत्र उल्टातर्क देताहै; कि grading अकेलाकारण है; कि hallucination पूरीतरह खत्म कीजासकतीहै; याकि case study चार मॉडलोंकीनिष्पक्ष ranking है।

मुख्य सीमाएँ: सही/गलत/"मुझे नहींपता" वालाजानबूझकर सरल मॉडल, जिसे लेखक "false trichotomy" कहते हैं; दस evaluations वालाछोटाcurated benchmark survey; uncontrolled case study — चार मॉडल, एक mitigation, एक test, default settings; और model evaluation कैसे होइस पर OpenAI से जुड़े लेखकोंकातर्क।

सामान्य पाठक कोकितनाभरोसारखनाचाहिए? इस बात पर उच्च कि hallucination न रहस्यमय है, न सख्त अर्थ में पूरी तरह अनिवार्य, और वर्तमान mainstream benchmarks अक्सर अनुमान लगाने कोपुरस्कृत करते हैं। प्रस्तावित सुधार बड़े पैमाने पर कितनामदद करेगा, इस पर मध्यम भरोसारखें — वास्तविक proof of concept है, व्यापक प्रदर्शन अभी नहीं।

स्रोत

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`