The CLEAN PAPER

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

ماڈلز guess کیوں کرتے ہیں — اور ہم نے انہیں guess کرنا کیوں سکھایا

زبان-ماڈل *confident* غلط *answers* دو *sources* سے آتے ہیں۔ شماریاتی: *patternless* حقائق ڈیٹا میں نایاب/*unseen* ہوں تو ماڈل کبھی غلط ہوگا، *singleton floor* شرح جیسے تخمینہ سے۔ *almost incentive*: ہر بینچ مارک "*I don't know*" کو غلط جیسا اسکور کرتا ہے، *guessing so* حتیٰ—*wins* کہ زیادہ غلط ماڈل زیادہ دیانت دار *abstaining* ماڈل سے اسکور کر سکتا ہے۔ مصنفین کھلا *propose rubrics* کرتے ہیں: اسکورنگ سوال میں حالت کریں۔ *frontier four* ماڈلز کے *limited* کیس مطالعہ میں یہ *flip incentive* کرتا ہے اور *abstention hallucination-reducing* طریقہ انعام ہوتی ہے۔ نظریہ+سروے + چھوٹا *uncontrolled* تجربہ، *promising* درست کرنا، وسیع *efficacy deployed* ابھی ثابت کرنا نہیں۔ *hallucinations* نہ *mystical* ہیں، نہ *strictly* ۔*inevitable*

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 1050–1047 (2026) · DOI: w-10549-026-s41586/1038.10

ماڈل اندازہ کیوں لگاتے ہیں — اور ہم نے انہیں ایسا کرنا کیوں سکھایا

کسی بڑے زبان ماڈل سے کسی غیر معروف شخص کی تاریخ پیدائش پوچھیں تو وہ ''7 مارچ'' پورے اعتماد سے بتا سکتا ہے، جیسے کسی ریکارڈ سے پڑھ رہا ہو — اور غلط ہو سکتا ہے، بلکہ دوبارہ پوچھنے پر تین مختلف تاریخیں بھی دے سکتا ہے۔ مصنفین اسی قسم کی مثالیں دیتے ہیں: سادہ حقائق کے سوال، مثلاً تاریخ پیدائش یا کوئی غیر معروف مخفف، جن پر ماڈل اعتماد کے ساتھ الگ الگ جواب گھڑ لیتے ہیں اور کوئی بھی درست نہیں ہوتا۔ صنعت اسے *hallucination* کہتی ہے، گویا یہ ادراک کی کوئی پراسرار خرابی ہو۔ مقالے کا پہلا قدم یہی پراسراریت ختم کرنا ہے۔

ماڈل کی سب سے بڑی ابتدائی تربیت، یعنی پری ٹریننگ، میں وہ بہت بڑے متنی ذخیرے سے زبان کے نمونے سیکھتا ہے۔ اب کسی ایسی حقیقت کو لیجیے جس کے پیچھے کوئی قابلِ اخذ نمونہ نہیں، مثلاً کسی خاص شخص کی تاریخ پیدائش۔ اگر وہ تاریخ تربیتی متن میں صرف ایک بار آئی ہو یا کبھی نہ آئی ہو تو نمونہ سیکھنے والے نظام کے پاس اسے اخذ کرنے کے لیے کافی شواہد نہیں ہوتے۔ مصنفین Turing Alan سے وابستہ ایک پرانے شماریاتی خیال سے استدلال کرتے ہیں: اگر تاریخوں میں سے خاصا حصہ صرف ایک بار دیکھا گیا ہو تو نئے، پہلے نہ دیکھے گئے حقائق پر غلطی کی ایک کم از کم سطح ناگزیر ہوگی۔ اس کے برعکس ممالک کے دارالحکومت جیسے حقائق متن میں بار بار دہرائے جاتے ہیں، اس لیے ان پر غلطی بہت کم ہوتی ہے۔ درست اور بظاہر قرینِ قیاس مگر غلط بیان میں فرق کرنا بھی خود ایک مشکل درجہ بندی کا مسئلہ ہے؛ صرف درست متن پیدا کرنا اس سے آسان نہیں۔ یعنی کچھ بنیادی غلطی واقعی پری ٹریننگ میں شامل ہے۔

بنیادی خیال: جو دیکھا ہی نہیں گیا، اسے کیسے گا جائے

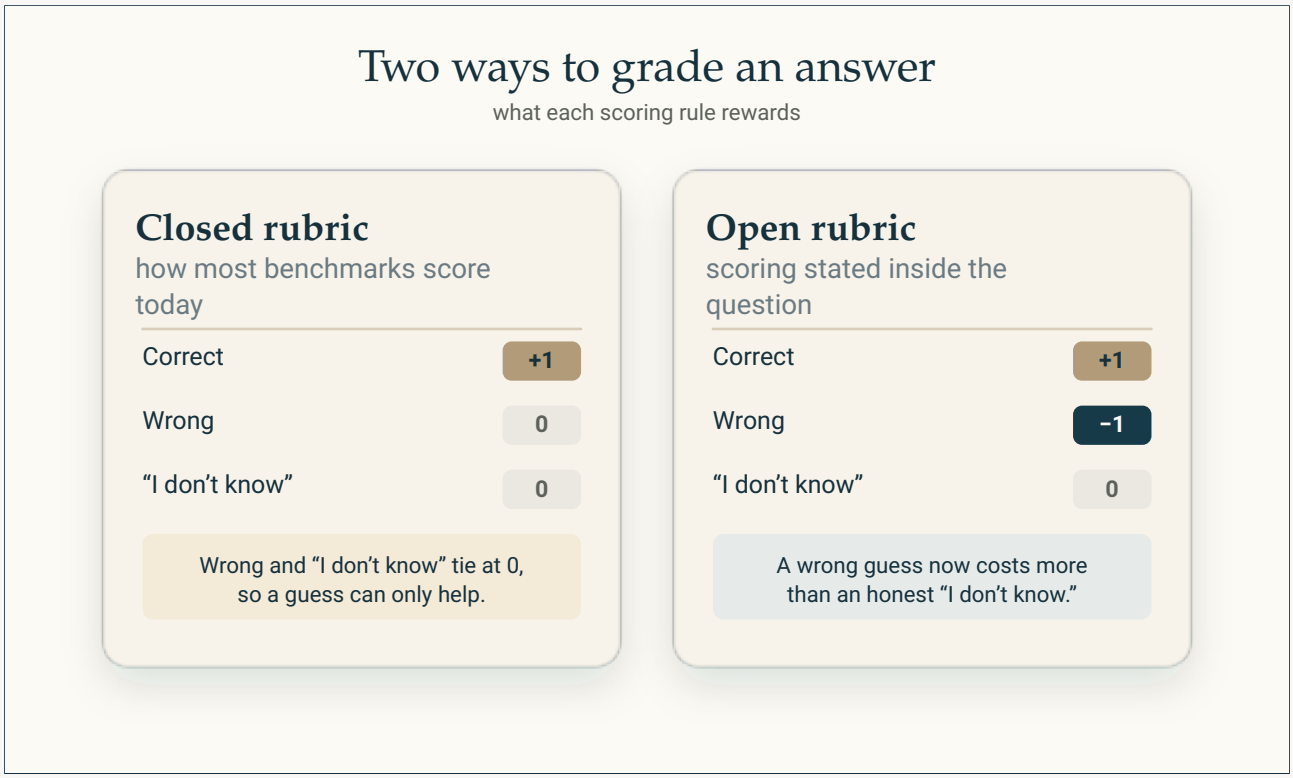
گیند بار 100 نہیں۔ معلوم فہرست مکمل کی رنگوں کو آپ مگر ہیں، گیندیں کی رنگوں مختلف میں تھیلے ایک کریں فرض اندازہ Good–Turing ہیں۔ دیتے دکھائی **بار ایک ایک صرف رنگ الگ 10** اور بار، 25 نیلی بار، 40 سرخ پر نکالنے نہیں راست براہِ رنگ دیکھے ان گیا۔ دیکھا نہیں کبھی پہلے جو ہے کتنا امکان کا آنے رنگ ایسا بار اگلی کہ ہے پوچھتا 10 میں 100 اگر ہیں۔ دیتے اشارہ ایک — ''سنگلٹن'' یعنی — رنگ والے دینے دکھائی بار ایک صرف مگر سکتے، جا گنے گئے۔ نہیں دیکھے ابھی جو ہے سکتا ہو میں رنگوں ایسے وزن احتمالی 10% اندازاً تو ہوں سنگلٹن مشاہدے

سنگلٹن خود غلطیاں نہیں، بلکہ لاعلمی کا پیمانہ ہیں: اگر بہت سی چیزیں صرف ایک بار نظر آئیں تو اس کا مطلب ہے کہ دنیا کا ایک قابلِ ذکر حصہ اب بھی نمونے سے باہر ہے۔

اب رنگوں کی جگہ تاریخ پیدائش اور تھیلے کی جگہ تربیتی متن رکھ دیں۔ اگر دیکھی گئی تاریخوں میں سے **ہر پانچ میں ایک صرف ایک بار** آئی ہو تو امکان کے تقریباً پانچویں حصے کا تعلق ایسے حقائق سے ہوگا جنہیں ماڈل نے قابلِ اعتماد طور پر نہیں سیکھا۔ تاریخ پیدائش حساب لگا کر اخذ نہیں کی جا سکتی؛ اگر کوئی حقیقت صرف ایک بار یا بالکل نہیں دیکھی گئی تو ماڈل کے پاس اس کے لیے مضبوط شماریاتی بنیاد نہیں ہوتی۔ زیادہ ذہانت بھی غائب ڈیٹا کو جادو سے پیدا نہیں کر سکتی۔ چھوٹے پیمانے پر یہی پوری دلیل ہے: سنگلٹن کی شرح بتاتی ہے کہ ڈیٹا سے دنیا کا کتنا حصہ ابھی قابلِ اعتماد طور پر نہیں سیکھا گیا، اور یہی غلطی کی ایک **کم از کم سطح** بناتی ہے۔ دارالحکومت جیسے بار بار دہرائے جانے والے حقائق میں سنگلٹن تقریباً صفر ہوتے ہیں، اس لیے وہ کہیں آسانی سے سیکھے جاتے ہیں۔

یہ اس بات کی وضاحت کرتا ہے کہ غلط جواب پیدا کہاں سے ہوتے ہیں، مگر یہ نہیں کہ دیانت داری اور مددگار ہونے کی بعد کی تربیت کے باوجود ماڈل ان پر اعتماد کیوں ظاہر کرتے رہتے ہیں۔ مصنفین امتحان دینے والے طالب علم کی مثال دیتے ہیں: خالی چھوڑنے پر صفر، اندازہ لگانے پر شاید ایک نمبر — تو زیادہ نمبر کے لیے اندازہ لگانا معقول حکمتِ عملی بن جاتا ہے۔ ماڈل بھی ایسا ہی حوصلہ افزا نظام دیکھتے ہیں، کیونکہ بیشتر بینچ مارک ''مجھے نہیں معلوم'' کو بھی غلط جواب کے برابر صفر دیتے ہیں۔ ایسی اسکورنگ میں ہر سوال کا جواب گھڑنے والا ماڈل، باقی سب لحاظ سے یکساں مگر غیر یقینی پر جواب روک دینے والے ماڈل سے

بہتر اسکور کر سکتا ہے۔ یعنی ہم خود اپنی پیمائشوں کے ذریعے اندازہ لگانے کی ترغیب دیتے ہیں۔



بینچ مارک اندازہ لگانے کو عقلی انتخاب بنا سکتے ہیں۔ عام اسکورنگ میں غلط جواب اور دیانت دار ''مجھے نہیں معلوم'' دونوں کو صفر ملتا ہے، اس لیے اندازہ لگانے میں صرف ممکنہ فائدہ ہے۔ اگر سوال ہی میں قواعد واضح کر دیے جائیں اور غلط جواب پر منفی نمبر ہوں تو غیر یقینی صورت میں جواب نہ دینا بہتر حکمتِ عملی بن سکتا ہے۔ یہ تبدیلی ترغیب کو بدلتی ہے؛ خود بخود تمام غلط جوابات ختم نہیں کرتی۔

اہم نکتہ عام سرخی کے برعکس ہے۔ ماڈل کی غلط بیانی کو اکثر ٹیکنالوجی کی کوئی پراسرار اور ناگزیر حد سمجھا جاتا ہے۔ مقالہ دونوں باتوں سے اختلاف کرتا ہے۔ پری ٹریننگ میں غلطی کی کم از کم سطح کوئی معما نہیں، بلکہ معمول کی شماریاتی غلطی ہے؛ اور بعد میں اس غلطی کا اعتماد کے ساتھ برقرار رہنا جزوی طور پر ان ترغیبات کا نتیجہ ہے جو ہم نے خود بنائی ہیں اور بدل سکتے ہیں۔ اگر کوئی نظام غیر یقینی ہونے پر سیدھا جواب دینے سے انکار کر دے تو وہ اعتماد سے غلط حقیقت نہیں گھڑے گا؛ تعینات ماڈل ایسا کم کرتے ہیں کیونکہ ہماری اسکورنگ اکثر انکار کو سزا دیتی ہے۔

مصنفین نے کیا کیا

مقالے کے تین حصے ہیں۔ پہلے حصے میں ریاضیاتی دلیل دی گئی ہے کہ پری ٹریننگ کے دوران کچھ غلط جوابات شماریاتی طور پر ناگزیر ہیں، اور ''صرف درست متن پیدا کرنا'' کم از کم اتنا ہی مشکل ہے جتنا یہ درجہ بندی کرنا کہ کوئی بیان درست ہے یا نہیں۔ دوسرے حصے میں دس بااثر بینچ مارکس کا جائزہ لیا گیا، جہاں مصنفین دکھاتے ہیں کہ عام درستگی کے پیمانے اندازہ لگانے کو انعام دیتے ہیں۔ تیسرے حصے میں ایک تجویز اور چھوٹا کیس مطالعہ ہے: **واضح اسکورنگ قواعد**، مثلاً سوال ہی میں یہ بتانا کہ درست جواب +1، غلط جواب −1، اور 50% سے کم یقین پر جواب نہ دینا بہتر ہے۔ پھر Gemini 3 Pro، GPT-5 Grok 4 اور Claude Opus 5.4 کو SimpleQA کے 4,326 حقائقی سوالات پر آزماتا گیا۔ مصنفین خود واضح کرتے ہیں کہ یہ محض وضاحتی مثال ہے، **کنٹرول شدہ بین-ماڈل موازنہ نہیں**۔

انہوں نے کیا پایا

- **پری ٹریننگ کچھ غلطی ناگزیر بناتی ہے۔** اعتماد کے ساتھ دیے گئے غلط جواب کی شرح کو بہترین درستگی-جانچنے والے درجہ بند کی غلطی سے جوڑا جا سکتا ہے؛ ایسے بے نمونہ حقائق میں جہاں معلومات بہت کم ہوں، سنگلٹن کی شرح غلطی کی ایک عملی کم از کم حد دیتی ہے۔ صاف ڈیٹا بھی ہر غلط جواب ختم نہیں کر سکتا۔
- **اسکورنگ واقعی اندازہ لگانے کو انعام دیتی ہے۔** عام درست/غلط اسکورنگ میں ہر سوال کا جواب دینا فائدہ مند رہتا ہے۔ جائزہ لیے گئے مقبول بینچ مارکس میں ''مجھے نہیں معلوم'' کو عموماً غلط جواب ہی سمجھا گیا۔ SimpleQA کی مثال میں خام درستگی قدرے **o4-mini** کے حق میں جاتی ہے، حالانکہ وہ تقریباً ہر سوال کا جواب دیتا ہے اور تین چوتھائی سے زیادہ پر غلط ہوتا ہے، جبکہ GPT-5-mini زیادہ بار جواب روک کر کم غلطیاں کرتا ہے۔ یعنی لاپرواہ ماڈل اسکور بورڈ پر بہتر دکھائی دے سکتا ہے۔
- **واضح قواعد کیس مطالعے میں ترغیب الٹ دیتے ہیں۔** ایک سادہ طریقہ یہ تھا کہ ماڈل سے دو بار جواب لیا جائے اور دونوں جواب مختلف ہوں تو جواب روک دیا جائے۔ روایتی درستگی میں اس سے غلطیاں کم ہوتی ہیں مگر درست جواب بھی کم ہو جاتے ہیں، اس لیے پیمانہ اسے سزا دیتا ہے۔ واضح منفی اسکور والے قاعدے میں یہی طریقہ چاروں ماڈلز کے لیے فائدہ مند ہو گیا؛ GPT-5-mini بھی o4-mini سے آگے نکلا جب اسکورنگ کا مقصد صاف بتایا گیا۔ ہر ماڈل کے لیے 4,326 سوالات استعمال ہوئے۔

غالب مطلب کیا ہے

غلط جوابات کم کرنے کے لیے شاید نئے، مخصوص ''ہیلوسینیشن ٹیسٹ'' بنانے سے زیادہ اہم یہ ہو کہ عام بینچ مارکس غیر یقینی کو کیسے اسکور کرتے ہیں۔ اگر اسکور بورڈ نہ بدلے تو جواب روکنے اور بہتر اعتماد-اندازے جیسے رویے خام درستگی کے پوائنٹس کھوتے رہیں گے۔ مصنفین اسے سماجی-تکنیکی مسئلہ سمجھتے ہیں: صرف ماڈل نہیں، پیمائش اور لیڈر بورڈ بھی رویہ بناتے ہیں۔

یہ کیا ثابت نہیں کرتا

- واضح اسکورنگ قواعد عملی دنیا میں غلط جواب ختم کر دیں گے — **نہیں۔** حمایتی تجربہ چھوٹا، جان بوجھ کر **غیر کنٹرول شدہ،** چار ماڈلز، ایک طریقہ اور ایک سوال جواب ٹیسٹ تک محدود تھا؛ اس نے صرف ترغیب بدلنے کا اصول دکھایا۔
- اسکورنگ **واحد سبب** ہے — نہیں۔ تربیتی ڈیٹا کی غلطیاں، مشکل مسائل اور غیر مانوس سوالات بھی الگ ذرائع ہیں۔
- غلط جواب ناگزیر ہی ہیں — مقالہ اس سے **اختلاف** کرتا ہے۔ ایسا نظام جو صرف قابلِ جانچ جواب دے اور باقی صورتوں میں ''مجھے نہیں معلوم'' کہے، اعتماد سے جھوٹی حقیقت گھڑنے سے بچ سکتا ہے، اگرچہ اس کی افادیت کم ہو سکتی ہے۔
- پری ٹریننگ کی کم از کم غلطی ختم ہو جاتی ہے — نہیں۔ مقالہ اسے سمجھاتا اور حد بندی کرتا ہے؛ ہر قسم کی حقائقی غلطی اسی ایک وجہ سے نہیں آتی۔
- واضح قواعد کافی ہیں — نہیں۔ بازیافت، بیرونی اوزار اور اعتماد کی بہتر پیمائش کی ضرورت اپنی جگہ رہتی ہے۔

شواھد کتنے مضبوط ھیں؟

بنیادی حصہ **ریاضیاتی** ھے — رسمی نچلی حدیں اور نظری دلائل، نہ کہ صرف تجرباتی پیمائشیں۔ اپنے مفروضوں کے اندر یہ مضبوط حصہ ھے۔

لیکن ماڈل جان بوجھ کر سادہ کیے گئے ھیں۔ مصنفین خود درست/غلط/''مجھے نہیں معلوم'' کی تین خانوں والی تقسیم کو ایک تقریب مانتے ھیں، اور بے نمونہ حقائق کی مثال بھی مثالی ھے۔ بینچ مارک سروے صرف دس منتخب ٹیسٹوں پر مشتمل ھے۔ کیس مطالعہ حقیقی ھے، مگر محدود اور غیر کنٹرول شدہ۔

ایک اور سیاق بھی اھم ھے: چار مصنفین میں سے تین موجودہ یا سابق **OpenAI** ملازم ھیں۔ اس لیے بینچ مارکنگ حکمتِ عملی بدلنے کی دلیل ایک ایسے فریق کی طرف سے آتی ھے جس کا میدان میں براہِ راست مفاد ھے؛ یہ خود دلیل کو رد نہیں کرتا، مگر آزاد بیرونی جانچ کی اھمیت بڑھاتا ھے۔ قابلِ ذکر بات یہ ھے کہ تنقید میں ان کے اپنے o4-mini اور GPT-5-mini کی مثالیں بھی شامل ھیں۔

یہ کیوں اھم ھے

یہ ایک بہت زیادہ مبالغہ آمیز مسئلے کو زیادہ معمولی اور قابلِ فہم شکل دیتا ھے۔ ''ھیلوسینیشن'' نہ صرف کوئی خوفناک پراسرار خرابی ھے، نہ لازماً ایک اٹل دیوار: اس میں ایک قابلِ فہم شماریاتی کم از کم سطح ھے اور اس کے اوپر ایسی ترغیبات ھیں جو ھم نے خود بنائی ھیں۔ وسیع تر سبق زیادہ خاموش مگر زیادہ مفید ھے: قابلِ اعتماد نظام بنانے میں پیش رفت شاید اتنی ھی اس بات پر منحصر ھو کہ ھم کیا ناپتے ھیں جتنی اس پر کہ ھم کیا بناتے ھیں۔

صاف خلاصہ

زبان ماڈلز کے اعتماد کے ساتھ دیے گئے غلط جواب دو بڑے ذرائع سے آ سکتے ھیں۔ پہلا شماریاتی ھے: ایسے بے نمونہ حقائق جو تربیتی ڈیٹا میں نایاب یا غائب ھوں، ان پر کچھ غلطی ناگزیر ھے اور سنگلٹن کی شرح جیسی مقدار اس کی کم از کم سطح کا اندازہ دے سکتی ھے۔ دوسرا ترغیبی ھے: بیشتر بینچ مارک ''مجھے نہیں معلوم'' کو بھی غلط جواب کے برابر اسکور کرتے ھیں، اس لیے اندازہ لگانا فائدہ مند ھو جاتا ھے — حتیٰ کہ زیادہ غلطیاں کرنے والا ماڈل زیادہ دیانت دار، جواب روکنے والے ماڈل سے بہتر اسکور کر سکتا ھے۔ مصنفین واضح اسکورنگ قواعد تجویز کرتے ھیں اور چار نمایاں ماڈلز کے ایک محدود کیس مطالعے میں دکھاتے ھیں کہ اس سے ترغیب واقعی بدل سکتی ھے۔ یہ نظریہ، سروے اور ایک چھوٹا غیر کنٹرول شدہ تجربہ ھے: دلچسپ اصلاحی سمت ضرور، مگر وسیع عملی اثر ابھی ثابت نہیں۔ غلط جواب نہ پراسرار ھیں اور نہ ھر صورت میں ناگزیر۔

بلا مبالغہ جانچ

مقالے کیا دکھاتا ھے: پری ٹریننگ میں غلطی کی ایک ریاضیاتی کم از کم سطح؛ جائزہ لیے گئے بیشتر مقبول بینچ مارکس میں جواب روکنے کا کوئی انعام نہیں؛ اور چار ماڈلز کے کیس مطالعے میں واضح اسکورنگ نے محتاط طریقے کو فائدہ پہنچایا۔

قرینِ قیاس مگر ثابت نہیں: عام بینچ مارکس میں واضح اسکورنگ قواعد اپنانے سے تعینات ماڈلز کے غلط جوابات معنی خیز حد تک کم ہوں گے۔

نہیں دکھاتا: کہ غلط جواب ناگزیر ہیں؛ کہ اسکورنگ واحد سبب ہے؛ کہ یہ طریقہ انہیں ختم کر دے گا؛ یا یہ کیس مطالعہ ماڈلز کی جامع درجہ بندی ہے۔

حدود: درست/غلط/نامعلوم کی سادہ تین خانوں والی تقسیم؛ بے نمونہ حقائق کا مثالی ماڈل؛ دس بینچ مارکس کا منتخب سروے؛ چار ماڈلز اور ایک ٹیسٹ پر غیر کنٹرول شدہ کیس مطالعہ؛ اور OpenAI سے وابستہ مصنفین کی بینچ مارکنگ حکمتِ عملی پر دلیل۔

اعتماد: اس بات پر بلند کہ موجودہ بینچ مارکس اندازہ لگانے کی ترغیب دیتے ہیں اور یہ مسئلہ کوئی شماریاتی معما نہیں۔ مجوزہ اصلاح کے وسیع پیمانے کے عملی اثر پر صرف درمیانہ اعتماد۔

ماخذ

(2026) 1050–1047 ,653 Nature
https://doi.org/10.1038/s41586-026-10549-w

(04664.2509 (arXiv Hallucinate Models Language Why
https://arxiv.org/abs/2509.04664

(OpenAI) hallucinations-paper-experiments
https://github.com/openai/hallucinations-paper-experiments

SimpleQA
https://github.com/openai/simple-evals