The CLEAN PAPER

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

چرا مدل‌های زبانی توهم می‌زنند — و چرا شیوهٔ نمره‌دهی ما آن را نگه می‌دارد

دروغ‌های مطمئنی که «توهم» می‌نامیم نقصی رازآلود نیستند: بخشی از آن‌ها محصول جانبی آماری آموزش‌اند، و باقی می‌مانند چون هایbenchmark اصلی حدس مطمئن را از «نمی‌دانم» صادقانه بیشتر پاداش می‌دهند. یک مطالعهٔ موردی روی چهار مدل مرزی نشان می‌دهد که بیان قواعد نمره‌دهی در *open prompt*» («*rubrics*» این انگیزه را وارونه می کند.

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

چرا مدل‌ها حدس می‌زنند، و چرا ما به آن‌ها چنین آموختیم

از یک مدل زبانی بزرگ تاریخ تولد یک غریبه را بپرسید؛ ممکن است با اعتمادبه‌نفس کسی که پاسخ را از روی کارت می‌خواند بگوید «۷ مارس» — و اشتباه کند، سه بار پشت سر هم، با سه تاریخ متفاوت. نویسندگان دقیقاً همین نوع مثال را می‌آورند: از مدل‌های پیشرو یک پرسش factual ساده پرسیده‌اند — تاریخ تولد یک شخص، یا معنی یک مخفف کم‌کاربرد — و هرکدام با اطمینان پاسخی متفاوت ساخته‌اند، بی‌آنکه هیچ کدام درست باشد. واژهٔ صنعت برای این پدیده توهم است، واژه‌ای که آن را شبیه نقصی در ادراک نشان می‌دهد. حرکت اول مقاله این است که رازآلودگی را از آن بگیرد.

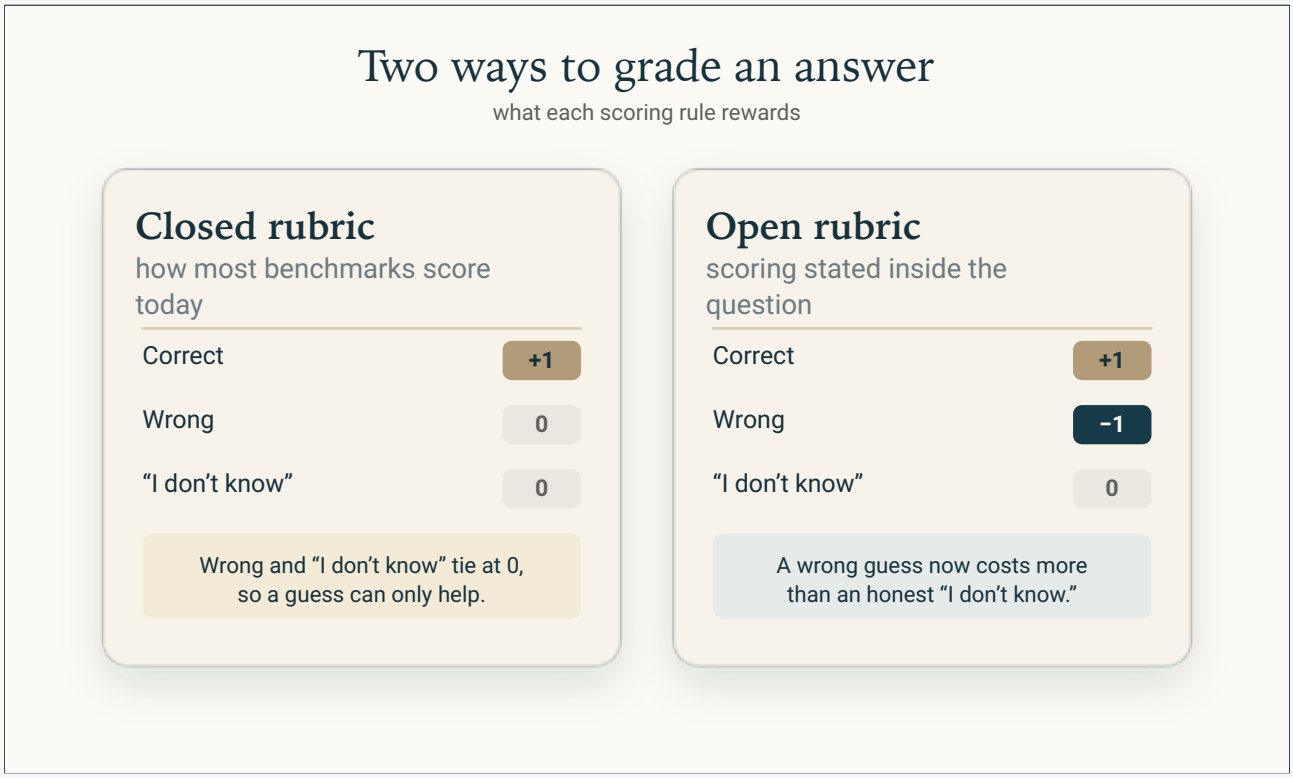
از شیوهٔ ساخته شدن مدل شروع کنید. در نخستین و بزرگ‌ترین مرحلهٔ آموزش، مدل در عمل می‌آموزد زبان روان چه شکلی دارد، با خواندن حجم عظیمی از متن. حالا یک واقعیت بی‌الگو را در نظر بگیرید — تاریخ تولد یک شخص خاص. اگر آن تاریخ در متن آموزش یک بار آمده باشد، یا اصلاً نیامده باشد، چیزی نیست که یک الگوآموز به آن چنگ بزند: پاسخ، از دید مدل، دل‌بخواهی است. نویسندگان این را با وام گرفتن از ایده‌ای قدیمی دقیق می‌کنند (ایدهٔ آلن تورینگ، از مسئله‌ای دیگر): اگر یکی از هر پنج تاریخ تولد فقط یک بار در داده‌ها ظاهر شود، باید انتظار داشت مدل دست‌کم در یکی از هر پنج مورد خطا کند — نه چون خراب است، بلکه چون هرگز چیزی برای یاد گرفتن وجود نداشته است. (با همین منطق، مدل‌ها تقریباً هیچ‌وقت پایتخت یک کشور را اشتباه نمی‌گویند: این‌ها مدام ظاهر می‌شوند.) آن‌ها با دقت استدلال می‌کنند که تشخیص یک گزارهٔ درست از یک گزارهٔ غلطِ محتمل خودش مسئله‌ای سخت است، و تولید فقط گزاره‌های درست دست‌کم به همان سختی است. یک کف خطا در مسئله پخته شده است.

yet seen haven’t you what count to how underneath: idea The

شویم. آشنا آن با درست که دارد را آن ارزش و زبانی، مدل‌های از قدیمی‌تر — دارد تکیه هوشمندانه واقعاً ایده‌ای بر این **با کیسه‌ای از توپ‌های رنگی شروع کنید.** نمی‌دانید چند رنگ در آن هست. ۱۰۰ توپ را یکی‌یکی بیرون می‌کشید و می‌شمارید: قرمز ۴۰، آبی ۲۵، سبز ۱۵، زرد ۵، بنفش ۳، نارنجی ۲ — و سپس **ده رنگ متفاوت که هرکدام دقیقاً یک بار ظاهر شده‌اند.** اکنون پرسشی که تورینگ واقعاً با آن روبه‌رو بود، در مسئله‌ای کاملاً متفاوت: احتمال اینکه توپ بعدی رنگی باشد که اصلاً ندیده‌اید چقدر است؟ نمی‌توانید چیزی را که هرگز نکشیده‌اید بشمارید — اما **می‌توانید** رنگ‌هایی را بشمارید که دقیقاً یک بار دیده‌اید، یعنی «تک‌نماها». ترفند، که **برآورد Good–Turing** نام دارد، این است که سهم برداشت‌هایی که تک‌نما بوده‌اند احتمال پنهان‌مانده در رنگ‌هایی را تخمین می‌زند که هنوز ندیده‌اید. ده تا از صد برداشت شما رنگ‌هایی بودند که فقط یک بار دیده شدند، پس احتمال اینکه توپ بعدی رنگی کاملاً تازه باشد حدود **10 / 100 = 10%.**
آن رنگ‌های یک‌باردیده اشتباه نیستند. آن‌ها اندازه‌گیری نادانی شما هستند: زیاد بودن رنگ‌هایی که فقط یک بار ظاهر می‌شوند راه نمونه برای گفتن این است که جهان چیزهای بیشتری دارد که شما هنوز نکشیده‌اید.
حالا رنگ‌ها را با تاریخ تولد عوض کنید، و کیسه را با متن آموزشی مدل. فرض کنید در میان تاریخ تولدهایی که دیده است، **یکی از هر پنج مورد دقیقاً یک بار ظاهر می‌شود.** همان ترفند: حدود یک پنجم احتمال در تاریخ تولدهایی زندگی می‌کند که مدل عملاً ندیده است — و تاریخ تولد الگویی ندارد که بتوان به آن تکیه کرد (نمی‌توانید تاریخ تولد کسی را محاسبه کنید). بنابراین تاریخی که یک بار دیده شده، یا هرگز دیده نشده، سکه‌ای است که مدل نمی‌تواند وزنش را تعیین کند، و در حدود **یکی از هر پنج** مورد اشتباه خواهد کرد. هیچ مقدار هوشمندی این را درست نمی‌کند: چیزی برای یادگیری وجود نداشت.
این کل استدلال در مقیاس کوچک است: نرخ تک‌نماها اندازه می‌گیرد چه مقدار از جهان از این داده‌ها یادگرفتنی نیست، و همان به **کفی** زیر خطاها تبدیل می‌شود. به همین دلیل هم مدل تقریباً هیچ‌وقت یک پایتخت را جا نمی‌اندازد — *Paris* مدام ظاهر می‌شود، نرخ تک‌نمای آن نزدیک صفر است، پس چیز زیادی برای یاد گرفتن وجود دارد.

این توضیح می‌دهد توهم‌ها از کجا می‌آیند. اما توضیح نمی‌دهد چرا باقی می‌مانند — چرا مدل‌ها، پس از همهٔ آموزش‌های بعدی که قرار بوده مفید و راستگو شوند، همچنان بلوف می‌زنند به جای آنکه تردید را بپذیرند. اینجا قیاس مقاله تقریباً آزاردهنده‌قدر مناسب است. دانش‌آموزی

را تصور کنید که در امتحان پاسخ را نمی‌داند. اگر پاسخ خالی صفر بگیرد و حدس ممکن است یک امتیاز بگیرد، حرکت بیشینه‌کنندهٔ نمره حدس زدن است — با اعتمادبه‌نفس، مشخص، و هرگز «مطمئن نیستم». دانش‌آموزان این را یاد می‌گیرند. ظاهراً مدل‌ها هم همین‌طور — چون ما به همان شکل نمره‌شان می‌دهیم. نویسندگان معیارهایی را که میدان واقعاً بر سر آن‌ها رقابت می‌کند، هایleaderboard را که مدل‌ها برای بالا رفتن از آن‌ها تنظیم می‌شوند، بررسی کرده‌اند و دیده‌اند تقریباً همهٔ آن‌ها به «نمی‌دانم» دقیقاً همان نمرهٔ پاسخ غلط را می‌دهند: صفر. تحت چنین قاعده‌ای، مدلی که همیشه حدس می‌زند از مدلی که جز این یکسان است اما صادقانه عدم قطعیتش را علامت می‌زند جلو می‌افتد. ما، به معنایی نسبتاً لفظی، با نمره‌دهی آن‌ها را به این سو می‌رانیم.



معیارها می‌توانند حدس زدن را عقلانی کنند. در نمره‌دهی‌ای که بیشتر معیارها استفاده می‌کنند (چپ)، پاسخ غلط و «نمی‌دانم» صادقانه هر دو صفر می‌گیرند — پس حدس فقط می‌تواند کمک کند. اگر قواعد در خود سؤال بیان شود، با جریمه برای خطا (راست)، پرهیز از پاسخ هنگام تردید می‌تواند حرکت بهتر شود. این عوض می‌کند آزمون چه چیزی را پاداش می‌دهد؛ به‌تنهایی توهم را حل نمی‌کند.

این بخشی است که باید نگه داشت، چون خلاف تیتر معمول حرکت می‌کند. توهم اغلب به‌عنوان حدی ناگزیر و تقریباً اسرارآمیز برای فناوری فروخته می‌شود. مقاله در هر دو مورد مخالفت می‌کند. کف پیش‌آموزش راز نیست — خطای آماری عادی است، از نوعی که یادگیری ماشین دهه‌هاست می‌فهمد. و تداوم آن ناگزیر نیست — بخشی از آن انگیزه‌ای است که خودمان ساخته‌ایم و می‌توانیم تغییر دهیم. سامانه‌ای که وقتی مطمئن نیست صرفاً از پاسخ دادن سر باز زند اصلاً توهم نخواهد داشت؛ دلیل اینکه مدل‌های مستقر چنین رفتار نمی‌کنند این است که هایscoreboard ما امتناع را تنبیه می‌کنند.

نویسندگان چه کردند

مقاله سه بخش دارد. نخست، استدلالی ریاضی که مقداری توهم در پیش‌آموزش از نظر آماری ناگزیر است، با نشان دادن اینکه «تولید فقط متن معتبر» دست‌کم به سختی یک مسئلهٔ طبقه‌بندی دودویی «آیا این گزاره معتبر است؟» است. دوم، استدلالی — با پشتیبانی یک پیمایش از ده benchmark اثرگذار — که هایmetric سبک دقتِ رایج حدس زدن را برپرهیز ترجیح می‌دهند. سوم، اصلاح پیشنهادی و یک مطالعهٔ موردی: ارزیابی‌های **open-rubric**، که در آن قواعد نمره‌دهی داخل خود سؤال بیان می‌شود (مثلاً «پاسخ درست ۱ می‌گیرد، پاسخ غلط −۱، پس اگر کمتر از ٪۵۰ مطمئن هستید خودداری کنید»)، تا مدل بداند چه وقت صداقت پاداش می‌گیرد. آن‌ها این را روی چهار مدل مرزی — Google's Gemini 3 Pro، OpenAI's GPT-5، xAI's Grok 4 وAnthropic's Claude Opus 5.4 — با 4,326 پرسش factual در SimpleQA آزموده‌اند. صریح می‌گویند مطالعهٔ موردی illustrative است، «نه ارزیابی کنترل‌شدهٔ بین مدل‌ها» (تنظیمات پیش‌فرض، بدون tuning، بدون نرمال‌سازی هزینه).

چه یافتند

- **پیش‌آموزش مقداری خطا را تحمیل می‌کند.** نرخ تولید دروغ‌های مطمئن از سوی مدل از پایین با تقریباً دو برابر نرخ خطای بهترین طبقه‌بند «آیا این گزاره معتبر است؟» ساخته‌شده از آن محدود می‌شود. برای facts بدون الگوی یادگرفتنی، آن کف دست‌کم نرخ تک‌نما است — سهم یfacts که دقیقاً یک بار در آموزش آمده‌اند. حتی با دادهٔ کاملاً پاک هم مقداری توهم اجتناب‌ناپذیر است.
- **نمره‌دهی به‌طور ملموس حدس را پاداش می‌دهد.** تحت نمره‌دهی معمول درست/نادرست، هرگز خودداری نکردن راهبرد بهینه است، و پیمایش نویسندگان نشان می‌دهد اکثریت بزرگ هایbenchmark محبوب «نمی‌دانم» را صرفاً غلط حساب می‌کنند. نمونه‌ای روشن از دادهٔ خودشان: در آزمون SimpleQA، دقت خام کمی به نفع OpenAI's o4-mini است — مدلی که تقریباً به همه چیز پاسخ می‌دهد و بیش از سه چهارم مواقع غلط است — نسبت به GPT-5-mini، که چون هنگام تردید خودداری می‌کند خطاهای بسیار کمتری دارد. مدل بی‌احتیاط‌تر روی scoreboard بهتر دیده می‌شود.
- **rubrics open انگیزه را وارونه می‌کنند (در مطالعهٔ موردی آن‌ها).** آن‌ها یک mitigation ساده برای توهم را آزمودند (مدل دو بار پاسخ دهد و اگر دو پاسخ ناسازگار بود خودداری کند). تحت دقت استاندارد، mitigation خطاها را کم می‌کند اما دقت را هم کم می‌کند — پس metric از به‌کارگیری آن بازمی‌دارد. تحت open rubrics، همان mitigation برای هر چهار مدل در بازه‌ای از جریمه‌ها بهتر بیرون می‌آید، و GPT-5-mini — که دقت خام به خاطر خودداری هنگام تردید تنبیش کرده بود — وقتی نمره‌دهی آشکار بیان شود از o4-mini جلو می‌افتد (n = 4,326 پرسش برای هر مدل).

این احتمالاً چه معنایی دارد

کاهش توهم عمدتاً مسئلهٔ اختراع آزمون‌های اختصاصی بیشتر برای توهم نیست. مسئله تغییر نحوهٔ نمره‌دهی هایbenchmark اصلی به عدم قطعیت است، تا اعتراف به «نمی‌دانم» دیگر تنبیه نشود. تا وقتی scoreboard عوض نشود، کاهش توهم همچنان برای مدل‌ها امتیاز دقت هزینه خواهد داشت و بنابراین همچنان بازداشته خواهد شد — به همین دلیل نویسندگان مسئله را «اجتماعی-فنی» می‌نامند: بخشی metric بهتر، بخشی واداشتن هایleaderboard اثرگذار به پذیرش آن.

این چه چیزی را ثابت نمی‌کند

- نشان نمی‌دهد rubrics open توهم را در عمل حل می‌کنند. آزمایش پشتیبان یک مطالعهٔ موردی کوچک و عمداً **کنترل‌نشده** است — چهار مدل با تنظیمات پیش‌فرض، یک mitigation انتخابی، یک آزمون factual-QA — برای نشان دادن وارونگی انگیزه، نه رتبه‌بندی مدل‌ها یا اثبات کارایی عمومی.
- ادعا نمی‌کند نمره‌دهی تنها علت است. خطاهای دادهٔ آموزش، مسائل واقعاً سخت، و هایprompt ناآشنا همچنان منابع جداگانه‌اند.
- از خط رایجِ اجتناب‌ناپذیر بودن توهم‌ها پشتیبانی نمی‌کند. نویسندگان برعکس استدلال می‌کنند: سامانه‌ای که فقط به پرسش‌های قابل بررسی پاسخ دهد و در غیر این صورت بگوید «نمی‌دانم» هرگز توهم نخواهد داشت.
- کف پیش‌آموزش را ناپدید نمی‌کند — آن را توضیح و کران‌گذاری می‌کند، و این کران مربوط به خطاهای factual مطمئن است، نه همهٔ رفتار مدل.
- نشان نمی‌دهد rubrics open به‌تنهایی کافی هستند. آن‌ها عوض می‌کنند ارزیابی چه چیزی را پاداش می‌دهد؛ جایگزین retrieval، استفاده از ابزار یا مدل‌های بهتر کالیبره‌شده نیستند.

شواهد چقدر قوی است؟

- هسته **ریاضی** است — کران‌های پایین رسمی، نه اندازه‌گیری. به‌عنوان استدلال نظری، در چارچوب خودش محکم است.
- بر **مدل‌های عمداً ساده‌شده** از مسئله تکیه دارد؛ خود نویسندگان «سه‌گانهٔ کاذب» درست/نادرست/«نمی‌دانم» و setting آرمانی facts» دل‌بخواهی» برای پاک‌ترین کران را علامت می‌زنند.
- پیمایش هاbenchmark **نمونه‌ای کوچک و گزینشی** است — ده ارزیابی اثرگذار، نه audit کامل.
- مطالعهٔ موردی **واقعی اما محدود** است: چهار مدل مرزی، یک mitigation، فقط SimpleQA، تنظیمات پیش‌فرض، صراحتاً «نه ارزیابی کنترل‌شده». این concept of proof برای استدلال انگیزه است، نه نتیجهٔ benchmark.
- ذکر point vantage مهم است: سه نفر از چهار نویسنده در **OpenAI** کار می‌کنند یا کرده‌اند، و مقاله استدلال می‌کند میدان باید نحوهٔ ارزیابی مدل‌ها را تغییر دهد. این موضعی خوب‌استدلال‌شده از طرف ذی‌نفع است، نه review بیرونی بی‌طرف — باید وزن شود، نه کنار گذاشته. (به نفع مقاله، نقد را به مدل‌های خود، o4-mini وGPT-5-mini، همان‌قدر راحت وارد می‌کند که به دیگران.)

چرا مهم است

این مقاله مسئله‌ای پرهیاهو را از نو قاب‌بندی می‌کند. «توهم» معمولاً یا عیب اسرارآمیز فروخته می‌شود یا دیواری جابه‌جانشدنی؛ این مقاله آن را عادی و تا حدی خودساخته می‌کند — کفی آماری که واقعاً می‌توانیم بفهمیم، روی انگیزه‌ای که خودمان انتخاب کرده‌ایم. درس گسترده‌تر آرام‌تر و مفیدتر است: پیشرفت بیشتر در قابلیت اعتماد شاید به همان اندازه به آنچه می‌سنجیم بستگی داشته باشد که به آنچه می‌سازیم.

خلاصهٔ تمیز

پاسخ‌های غلط اما مطمئن از مدل‌های زبانی از دو جا می‌آیند. اولی آماری است: وقتی یfact الگویی برای یادگیری ندارد، مدلی که برای تقلید زبان آموزش دیده گاهی آن را غلط می‌گوید، و آن کف را می‌توان تخمین زد (مثلاً از تعداد یfacts که فقط یک بار در آموزش ظاهر می‌شوند). دومی انگیزه است: تقریباً هر یbenchmark که مدل‌ها بر اساس آن رتبه‌بندی می‌شوند «نمی‌دانم» را مثل پاسخ غلط نمره می‌دهد، پس حدس زدن همیشه می‌برد — تا حدی که مدلی که سه چهارم مواقع غلط است می‌تواند از مدل صادق‌تری که خودداری می‌کند جلو بزند. پیشنهاد نویسندگان آزمون توهم دیگری نیست، بلکه open» «rubrics است: قواعد نمره‌دهی را داخل سؤال بگویید. در مطالعه‌ای موردی روی چهار مدل مرزی، این انگیزه را وارونه می‌کند تا روش کاهش‌دهندهٔ توهم پاداش بگیرد نه تنبیه. این مقاله‌ای نظری-همراه-پیمایش با آزمایشی کوچک و صریحاً کنترل‌نشده است، peer-reviewed در *Nature*؛ اصلاح امیدوارکننده است اما هنوز در مقیاس وسیع نشان داده نشده، و توهم‌ها نه اسرارآمیزند و نه به‌طور سخت‌گیرانه اجتناب‌ناپذیر.

بررسی بی‌تعارف

مقاله چه نشان می‌دهد: کران پایین ریاضی‌ای که طبق آن مقداری توهم در پیش‌آموزش تحمیل می‌شود (دست‌کم «نرخ تک‌نما» برای facts بی‌الگو)؛ پیمایشی که می‌گوید بیشتر هایbenchmark پیشرو به «نمی‌دانم» اعتبار نمی‌دهند؛ و مطالعه‌ای چهارمدلی که در آن بیان نمره‌دهی در open prompt») («rubrics روش کاهش‌دهندهٔ توهم را برنده می‌کند، جایی که دقت ساده آن را تنبیه کرده بود.

چه چیزی محتمل اما اثبات‌نشده است: اینکه open ،rubrics اگر وارد هایbenchmark اصلی شوند، توهم را در مدل‌های مستقر به شکل معنادار کاهش دهند. آزمایش پشتیبان کوچک و صریحاً کنترل‌نشده است.

چه چیزی را نشان نمی‌دهد: اینکه توهم‌ها اجتناب‌ناپذیرند (برعکس استدلال می‌کند)؛ اینکه نمره‌دهی تنها علت است؛ اینکه توهم را می‌توان کاملاً حذف کرد؛ یا اینکه مطالعهٔ موردی چهار مدل را در برابر هم رتبه‌بندی می‌کند.

محدودیت‌های اصلی: مدل ساده‌شدهٔ درست/نادرست/«نمی‌دانم» (خود نویسندگان آن را «سه‌گانهٔ کاذب» می‌نامند)؛ پیمایش کوچک و گزینشی هاbenchmark (ده ارزیابی)؛ مطالعهٔ موردی کنترل‌نشده (چهار مدل، یک ،mitigation یک آزمون، تنظیمات پیش‌فرض)؛ و استدلالی به رهبری OpenAI دربارهٔ اینکه میدان چگونه باید مدل‌ها را ارزیابی کند.

خوانندهٔ عمومی چقدر باید مطمئن باشد؟ زیاد، نسبت به اینکه توهم نه اسرارآمیز است و نه کاملاً اجتناب‌ناپذیر، و هایbenchmark اصلی اکنون حدس زدن را پاداش می‌دهند. متوسط، نسبت به اینکه اصلاح پیشنهادی کمک می‌کند — اکنون concept of proof واقعی دارد، اما هنوز نشان نداده در گستره و مقیاس وسیع کار می‌کند.

منابع

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