The CLEAN PAPER

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

لماذا تهلوس نماذج اللغة — ولماذا تُبقيها طريقة تقييمنا كذلك

الأكاذيب الواثقة التي نسميها «هلوسات» ليست خللًا غامضًا: بعضها نتيجة إحصائية جانبية للتدريب، وتستمر لأن المعايير الشائعة تكافئ التخمين الواثق على «لا أعرف» الصادقة. وتُظهر دراسة حالة على أربعة نماذج أمامية أن ذكر قواعد الدرجات في السؤال («معايير مفتوحة») يعكس هذا الحافز.

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

لماذا تخمّن النماذج، ولماذا علّمناها أن تفعل ذلك

اسأل نموذج لغة كبيرًا عن تاريخ ميلاد شخص غير معروف، وقد يجيبك: "7 مارس" بثقة من يقرأ الإجابة من بطاقة — ويكون مخطئًا، ثلاث مرات متتالية، بثلاثة تواريخ مختلفة. يعطي المؤلفون مثالًا من هذا النوع بالضبط: نماذج رائدة طُرحت عليها أسئلة واقعية بسيطة — تاريخ ميلاد شخص، أو معنى اختصار غامض — فاخترع كل واحد منها إجابة مختلفة بثقة، ولا واحدة صحيحة. الكلمة الشائعة في الصناعة لهذا هي هلوسة، وهي توحي بأن الخلل في الإدراك. خطوة الورقة الأولى هي نزع الغموض عنها.

ابدأ من طريقة بناء النموذج. في مرحلته التدريبية الأولى والأكبر، يتعلم عمليًا كيف تبدو اللغة الطليقة من خلال قراءة كمية هائلة من النصوص. خذ الآن حقيقة لا نمط وراءها — تاريخ ميلاد شخص بعينه. إذا ظهر ذلك التاريخ في نص التدريب مرة واحدة، أو لم يظهر أصلًا، فلا يوجد شيء يستطيع متعلم الأنماط أن يمسك به: الإجابة، من منظور النموذج، عشوائية. يجعل المؤلفون هذا دقيقًا باستعارة فكرة قديمة (فكرة آلان تورنغ، من مسألة أخرى): إذا كان واحد من كل خمسة تواريخ ميلاد لا يظهر في البيانات إلا مرة واحدة، فينبغي توقع أن يخطئ النموذج في واحد من كل خمسة منها على الأقل — ليس لأنه معطوب، بل لأنه لم يكن هناك شيء يتعلمه. (وبالمنطق نفسه، تكاد النماذج لا تخطئ في عاصمة بلد: فهذه تظهر باستمرار.) ويجادلون بحذر في أن تمييز العبارة الصحيحة من العبارة الخاطئة المعقولة هو نفسه مسألة صعبة، وأن إنتاج عبارات صحيحة فقط لا يقل صعوبة عن ذلك. هناك حد أدنى للخطأ مدمج في المسألة.

الفكرة الأساسية: كيف نعدّ ما لم نره بعد؟

جيدًا. تُفهم أن وتستحق اللغة، نماذج من أقدم — فعلًا ذكية حيلة إلى الفكرة تستند

ابدأ بكيس من كرات ملوّنة. أنت لا تعرف كم لونًا في داخله. تسحب 100 كرة، واحدة تلو الأخرى، وتعدّها: أحمر 40، أزرق 25، أخضر 15، أصفر 5، بنفسجي 3، برتقالي 2 — ثم **عشرة ألوان مختلفة ظهر كل واحد منها مرة واحدة فقط.**

السؤال الذي واجهه تورنغ فعليًا، في مسألة مختلفة تمامًا، هو: ما احتمال أن تكون الكرة التالية بلون لم تره قط؟ لا تستطيع أن تعد ما لم تسحبه أبدًا — لكنك **تستطيع** أن تعد الألوان التي رأيتها مرة واحدة فقط، أي "الأفراد". الحيلة، المسماة **تقدير Good–Turing**، هي أن حصة السحوبات التي كانت أفرادًا تقدّر الاحتمال الذي ما زال مختبئًا في الألوان التي لم ترها. عشرة من مئة سحبة كانت ألوانًا ظهرت مرة واحدة، إذن احتمال أن يكون اللون التالي جديدًا يقارب **10 / 100 = 10%.**

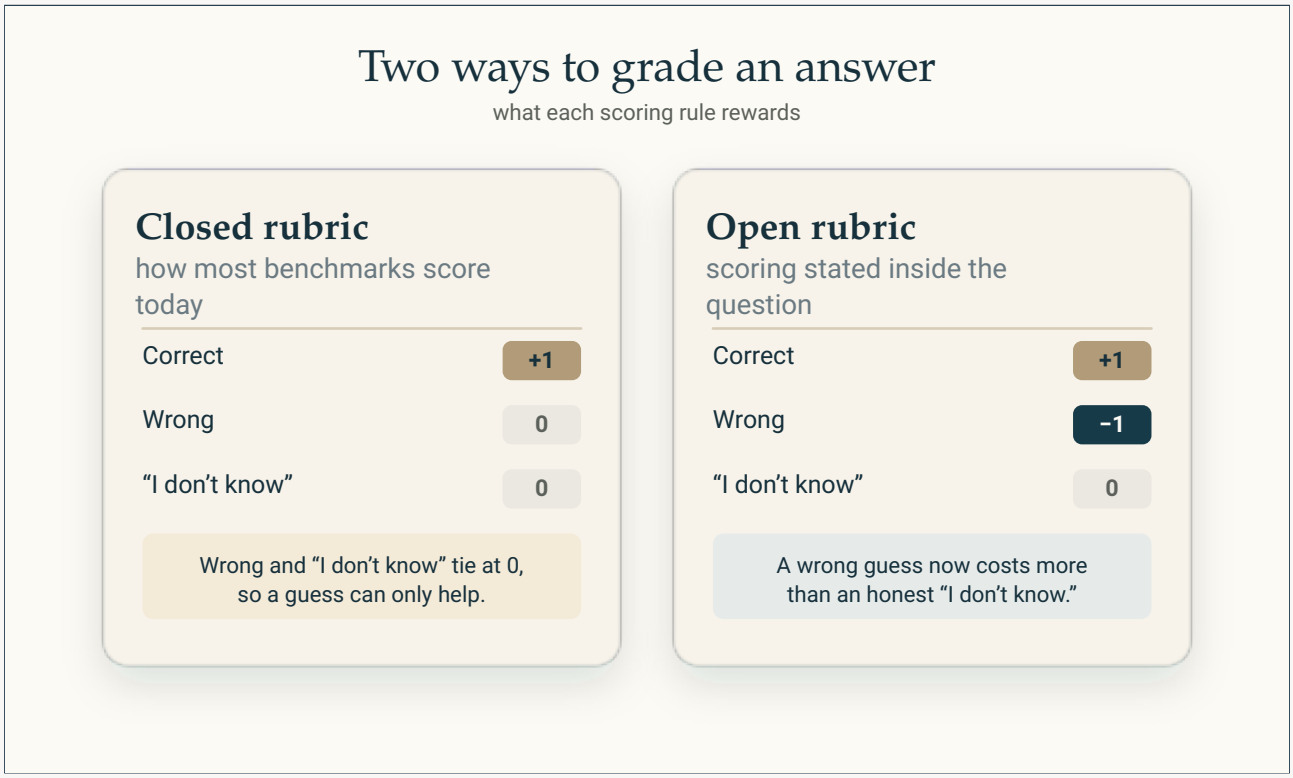
هذه الألوان ذات الظهور الواحد ليست أخطاء. إنها قياس لجهلك أنت: ظهور ألوان كثيرة مرة واحدة هو طريقة العينة في إخبارك بأن العالم يحتوي أكثر مما سحبته ببساطة.

استبدل الآن الألوان بتواريخ الميلاد، والكيس بنص تدريب النموذج. افترض أن، من بين تواريخ الميلاد التي رآها، **واحدًا من كل خمسة يظهر مرة واحدة فقط.** الحيلة نفسها: نحو خُمس الاحتمال يعيش في تواريخ ميلاد لم يرها النموذج فعليًا — وتاريخ الميلاد لا يملك نمطًا يُعوَّل عليه (لا يمكنك أن تستنتج تاريخ ميلاد شخص). لذلك فإن تاريخًا شوهد مرة واحدة، أو لم يُرَ، هو عملة لا يستطيع النموذج ترجيحها، وسيخطئ في نحو **واحد من كل خمسة** منها. لا مقدار من الذكاء يحل هذا: لم يكن هناك شيء ليتعلمه.

هذه هي الحجة كلها في صورة مصغرة: معدل الأفراد يقيس مقدار ما لا يمكن تعلمه من العالم من هذه البيانات، ويصبح ذلك **أرضية** تحت الأخطاء. ولهذا أيضًا يكاد النموذج لا يخطئ في مدينة عاصمة — باريس تظهر باستمرار، ومعدل الأفراد فيها قريب من الصفر، لذلك توجد مادة كثيرة للتعلم.

هذا يفسر من أين تأتي الهلوسات. لكنه لا يفسر لماذا تبقى — لماذا تستمر النماذج، بعد كل التدريب اللاحق المفترض أن يجعلها مفيدة وصادقة، في المخادعة بدل الاعتراف بالشك. هنا يكون تشبيه الورقة ملائمًا على نحو غير مريح. تخيل طالبًا في امتحان لا يعرف الإجابة. إذا كان ترك الفراغ يساوي صفرًا والتخمين قد يساوي نقطة، فالحركة التي تعظم الدرجة هي التخمين — بثقة، وبشكل محدد، ومن دون "لست متأكدًا". يتعلم الطلاب ذلك. ويبدو أن النماذج تتعلمه أيضًا — لأننا نقيّمها بالطريقة نفسها. راجع المؤلفون المعايير التي يتنافس عليها الحقل

فعليًا، ولوحات الصدارة التي تضبط النماذج كي تصعدها، ووجدوا أن معظمها تقريبًا يعطي "لا أعرف" الدرجة نفسها التي يعطيها للإجابة الخاطئة: صفرًا. تحت هذه القاعدة، سيتغلب نموذج يخمّن دائمًا على نموذج مطابق له إلا أنه يعلن عدم يقينه بصدق. نحن، بمعنى حرفي إلى حد بعيد، ندربها عبر نظام الدرجات على ذلك.



يمكن للمعايير أن تجعل التخمين عقلانيًا. تحت أسلوب الدرجات الذي تستخدمه معظم المعايير (يسارًا)، تحصل الإجابة الخاطئة و"لا أعرف" الصادقة على صفر — لذلك لا يمكن للتخمين إلا أن يساعد. إذا صيغت القواعد داخل السؤال، مع عقوبة على الخطأ (يمينًا)، فقد يصبح الامتناع عند الشك هو الخيار الأفضل. هذا يغيّر ما يكافئه الاختبار، ولا يحل، وحده، مشكلة الهلوسة.

هذا هو الجزء الذي يستحق الاحتفاظ به، لأنه يعاكس العنوان المعتاد. كثيرًا ما تُباع الهلوسة باعتبارها حدًا حتميًا، شبه غامض، للتقنية. تعارض الورقة ذلك في الجانبين. أرضية ما قبل التدريب ليست لغزًا — إنها خطأ إحصائي عادي، من النوع الذي يفهمه تعلم الآلة منذ عقود. والاستمرار ليس حتميًا — إنه، جزئيًا، حافز بنيناه ويمكننا تغييره. نظام يرفض ببساطة الإجابة عندما يكون غير واثق لن يهلوس إطلاقًا؛ سبب أن النماذج المنشورة لا تتصرف هكذا هو أن لوحات الدرجات تعاقب الرفض.

ماذا فعل المؤلفون

تتكون الورقة من ثلاثة أجزاء. أولًا، حجة رياضية بأن قدرًا من الهلوسة مفروض إحصائيًا أثناء ما قبل التدريب، عبر إظهار أن "توليد نص صالح فقط" لا يقل صعوبة عن مسألة تصنيف ثنائية: "هل هذه العبارة صالحة؟". ثانيًا، حجة — مدعومة بمسح لعشرة معايير مؤثرة — بأن مقاييس الدقة الشائعة تكافئ التخمين على الامتناع. ثالثًا، إصلاح مقترح ودراسة حالة تختبره: تقييمات **بمعيار مفتوح**، حيث تُذكر قواعد الدرجات داخل السؤال نفسه (مثلًا: "الإجابة الصحيحة تحصل على 1، والخاطئة −1، لذلك امتنع إذا كان يقينك أقل من 50%")، بحيث يستطيع النموذج معرفة متى تكافأ الصراحة. يختبرون ذلك على أربعة نماذج أمامية — GPT-5 OpenAI's وPro 3 Gemini Google's

وGrok xAI's 4 وClaude Anthropic's Opus 5.4 — باستخدام أسئلة SimpleQA الواقعية وعددها 4,326. وهم واضحون في أن دراسة الحالة توضيحية، "وليست تقييمًا مضبوطًا عبر النماذج" (إعدادات افتراضية، بلا ضبط، وبلا تسوية للتكلفة).

ماذا وجدوا

- **ما قبل التدريب يفرض بعض الخطأ.** معدل إصدار النموذج لعبارات خاطئة بثقة محدود من الأسفل بمعدل خطأ أفضل مصنف "هل هذه العبارة صالحة؟" مبني منه، تقريبًا مضروبًا في اثنين. وبالنسبة إلى حقائق بلا نمط قابل للتعلم، تكون هذه الأرضية على الأقل معدل الأفراد — أي حصة الحقائق التي تظهر مرة واحدة فقط في التدريب. بعض الهلوسة لا مفر منه حتى مع بيانات نظيفة تمامًا.
- **نظام الدرجات يكافئ التخمين عمليًا.** تحت أسلوب الصحيح/الخاطئ العادي، يكون عدم الامتناع أبدًا هو الاستراتيجية المثلى، ويجد مسح المؤلفين أن الغالبية الكبرى من المعايير الشعبية تسجل "لا أعرف" نكطأ ببساطة. مثال حي من طرفهم: في اختبار SimpleQA، تفضّل الدقة الخام قليلًا o4-mini OpenAI's — الذي يجيب تقريبًا عن كل شيء ويخطئ في أكثر من ثلاثة أرباع الحالات — على GPT-5-mini، الذي يرتكب أخطاء أقل بكثير لأنه يمتنع عندما لا يكون واثقًا. النموذج الأكثر تهورًا يبدو أفضل على لوحة الدرجات.
- **المعايير المفتوحة تقلب الحافز (في دراسة الحالة لديهم).** اختبروا وسيلة بسيطة لتقليل الهلوسة (أن يجيب النموذج مرتين ويمتنع إذا اختلفت الإجابتان). تحت الدقة القياسية، تقلل الوسيلة الأخطاء لكنها تقلل الدقة أيضًا — لذلك يثبط المقياس اعتمادها. تحت المعايير المفتوحة، تتقدم الوسيلة نفسها في النتيجة لكل النماذج الأربعة عبر طيف من العقوبات؛ كما أن GPT-5-mini — الذي عاقبته الدقة الخام لأنه امتنع عند الشك — يتقدم على o4-mini عندما تُذكر قواعد الدرجات صراحة (n = 4,326 سؤالًا لكل نموذج).

ماذا يعني هذا على الأرجح

تقليل الهلوسة ليس في الغالب مسألة اختراع المزيد من اختبارات الهلوسة المتخصصة. بل هو مسألة تغيير طريقة تسجيل المعايير الرئيسية لعدم اليقين، بحيث لا يعود الاعتراف بـ"لا أعرف" معاقبًا. إلى أن تتغير لوحة الدرجات، سيظل تقليل الهلوسة يكلف النماذج نقاط دقة، وبالتالي سيظل مثبطًا — ولهذا يصف المؤلفون المشكلة بأنها "اجتماعية-تقنية": جزء منها مقياس أفضل، وجزء منها دفع لوحات الصدارة المؤثرة إلى تبنيه.

ما الذي لا تثبته الورقة

- لا تُظهر أن المعايير المفتوحة تصلح الهلوسة في العالم الحقيقي. التجربة الداعمة دراسة حالة صغيرة **وغير مضبوطة** عمدًا — أربعة نماذج بإعدادات افتراضية، وسيلة واحدة مختارة، واختبار أسئلة واقعية واحد — هدفها إظهار قلب الحافز، لا ترتيب النماذج أو إثبات فعالية عامة.
- لا تدعي أن نظام الدرجات هو السبب الوحيد. أخطاء بيانات التدريب، والمسائل الصعبة فعلًا، والمطالبات غير المألوفة تظل مصادر منفصلة.
- لا تدعم العبارة الشعبية أن الهلوسات حتمية. يجادل المؤلفون بالعكس: نظام يجيب فقط عن الأسئلة القابلة للتحقق ويقول غير ذلك "لا أعرف" لن يهلوس أبدًا.
- لا تجعل أرضية ما قبل التدريب تختفي — بل تشرحها وتحدّها، ويتعلق الحد بالأخطاء الواقعية الواثقة، لا بكل سلوك النموذج.

- لا تُظهر أن المعايير المفتوحة كافية وحدها. هي تغيّر ما يكافئه التقييم؛ وليست بديلًا عن الاسترجاع أو استخدام الأدوات أو النماذج الأفضل معايرة.

ما مدى قوة الدليل؟

- اللب **رياضيات** — حدود دنيا شكلية، لا قياسات. كحجة نظرية، هي متماسكة وفق شروطها.
- تستند إلى **نماذج مبسطة عمدًا** للمشكلة؛ والمؤلفون أنفسهم يشيرون إلى "الثلاثية الزائفة" في معاملة كل إجابة كصحيحة أو خاطئة أو "لا أعرف"، وإلى إعداد "الحقائق العشوائية" المثالي المستخدم لأوضح حد.
- مسح المعايير **عينة صغيرة ومنتقاة** — عشرة تقييمات مؤثرة، لا تدقيق شامل.
- دراسة الحالة **حقيقية لكنها محدودة**: أربعة نماذج أمامية، وسيلة واحدة، SimpleQA فقط، إعدادات افتراضية، ومذكورة صراحة بأنها "ليست تقييمًا مضبوطًا". إنها إثبات مفهوم لحجة الحافز، لا نتيجة معيارية.
- يجدر ذكر زاوية النظر: ثلاثة من المؤلفين الأربعة يعملون أو عملوا في **OpenAI**، والورقة تجادل بأن الحقل ينبغي أن يغيّر طريقة تقييم النماذج. هذا موقف محكم من طرف ذي مصلحة، لا مراجعة خارجية محايدة — يُوزن، لا يُرفض. (ويحسب لها أن الورقة توجه النقد إلى نماذجها، o4-mini وGPT-5-mini، بالسهولة نفسها التي توجهه بها إلى غيرها.)

لماذا يهم

إنها تعيد تأطير مشكلة مثقلة بالضجيج. تميل "الهلوسة" إلى أن تُباع إما كعيب شبحي أو كجدار لا يتحرك؛ تجعلها هذه الورقة أمرًا عاديًا ومصنوعًا بأيدينا جزئيًا — أرضية إحصائية يمكننا فهمها فعلًا، تعلوها حوافز اخترناها. الدرس الأوسع أهدأ وأكثر فائدة: قد يعتمد مزيد من التقدم في الاعتمادية بقدر كبير على ما نقيسه كما يعتمد على ما نبنيه.

ملخص صافٍ

تأتي الإجابات الخاطئة الواثقة من نماذج اللغة من مصدرين. الأول إحصائي: عندما لا تملك الحقيقة نمطًا يمكن تعلمه، سيخطئ أحيانًا نموذج دُرّب على تقليد اللغة، ويمكن تقدير تلك الأرضية (مثلًا من عدد الحقائق التي تظهر مرة واحدة فقط في التدريب). والثاني هو الحوافز: كل معيار تقريبًا تُرتب النماذج بناء عليه يسجل "لا أعرف" مثل الإجابة الخاطئة، لذلك يفوز التخمين دائمًا — إلى حد أن نموذجًا يخطئ في ثلاثة أرباع الحالات يمكن أن يتفوق على نموذج أكثر صدقًا يمتنع. اقتراح المؤلفين ليس اختبار هلوسة آخر، بل "معايير مفتوحة": اذكر قواعد الدرجات داخل السؤال. في دراسة حالة على أربعة نماذج أمامية، يقلب ذلك الحافز بحيث تكافأ وسيلة تقلل الهلوسة بدل أن تعاقب. إنها ورقة نظرية مع مسح وتجربة صغيرة غير مضبوطة صراحة، خضعت لمراجعة الأقران في *Nature*؛ الإصلاح واعد لكنه لم يثبت بعد على نطاق واسع، والهلاوس ليست غامضة ولا حتمية تمامًا.

فحص بلا تجميل

ما تظهره الورقة: حد أدنى رياضي يفرض بعض الهلوسة أثناء ما قبل التدريب (على الأقل "معدل الأفراد" للحقائق بلا نمط)؛ ومسح يجد أن معظم المعايير الرائدة لا تعطي "لا أعرف" أي رصيد؛ ودراسة حالة بأربعة نماذج تجعل فيها قواعد الدرجات المعلنة داخل السؤال ("معايير مفتوحة") وسيلة تقليل الهلوسة رابحة، حيث كانت الدقة البسيطة تعاقبها.

ما هو معقول لكنه غير مثبت: أن المعايير المفتوحة، إذا أُدخلت في المعايير الرئيسية، ستقلل الهلوسة بقدر مهم في النماذج المنشورة. التجربة الداعمة صغيرة وغير مضبوطة صراحة.

ما لا تظهره: أن الهلوسات حتمية (هي تجادل بالعكس)؛ أن نظام الدرجات هو السبب الوحيد؛ أن الهلوسة يمكن القضاء عليها كليًا؛ أو أن دراسة الحالة ترتب النماذج الأربعة فيما بينها.

القيود الرئيسية: نموذج مبسط عمدًا لصحيح/خاطئ/"لا أعرف" (يسميه المؤلفون "ثلاثية زائفة")؛ مسح صغير ومنتقى للمعايير (عشرة تقييمات)؛ دراسة حالة غير مضبوطة (أربعة نماذج، وسيلة واحدة، اختبار واحد، إعدادات افتراضية)؛ وحجة تقودها OpenAI حول كيفية تقييم الحقل للنماذج.

ما مقدار الثقة المناسب للقارئ العام؟ ثقة عالية بأن الهلوسة ليست غامضة ولا حتمية تمامًا، وأن المعايير الرئيسية تكافئ التخمين حاليًا. ثقة متوسطة بأن الإصلاح المقترح يساعد — لديه الآن إثبات مفهوم حقيقي، لكن ليس بعد دليلًا على أنه يعمل على نطاق واسع.

المصادر

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