The Clean Paper

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

Почему языковые модели галлюцинируют — и почему наша система оценки помогает этому продолжаться

Уверенные ложные утверждения, которые мы называем «галлюцинациями», не являются загадочным сбоем: часть из них — статистический побочный эффект обучения, а сохраняются они еще и потому, что главные бенчмарки вознаграждают уверенную догадку сильнее честного «я не знаю». Case study на четырех frontier-моделях показывает, что явное формулирование правил оценки в самом запросе — «open rubrics» — переворачивает этот стимул.

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

Почему модели угадывают — и почему мы сами их этому научили

Спросите большую языковую модель о дне рождения незнакомого человека — и она может ответить «7 марта» со спокойной уверенностью того, кто читает дату с карточки, а затем три раза подряд ошибиться, каждый раз называя другую дату. Авторы приводят именно такие примеры: ведущим моделям задают простой фактический вопрос — день рождения человека или расшифровку малоизвестной аббревиатуры — и каждая уверенно придумывает свой ответ, причем ни один не верен. В области это называют *галлюцинацией*, как будто речь идет о сбое восприятия. Первый шаг статьи — убрать из этого явления мистику.

Начнем с того, как строят модель. На первом и крупнейшем этапе обучения она, по сути, учится тому, как выглядит естественный язык, читая огромный массив текста. Теперь возьмем факт без закономерности — конкретный день рождения определенного человека. Если эта дата встретилась в тренировочном тексте один раз или не встретилась вовсе, модели, обучающейся на паттернах, не за что зацепиться: с ее точки зрения ответ произволен. Авторы делают эту идею точной, используя старый подход Алана Тьюринга из другой задачи: если каждый пятый день рождения встречается в данных лишь один раз, следует ожидать, что модель будет ошибаться как минимум примерно в одном случае из пяти — не потому, что она «сломана», а потому, что в данных не было закономерности, которую можно выучить. (По той же логике модели почти никогда не ошибаются со столицами стран: эти факты повторяются постоянно.) Авторы аккуратно доказывают, что отличить истинное утверждение от правдоподобного ложного — само по себе сложная задача, а генерировать *только* истинные утверждения как минимум не проще. Значит, некоторый нижний предел ошибок заложен статистически.

Идея под этим результатом: как посчитать то, чего вы еще не видели

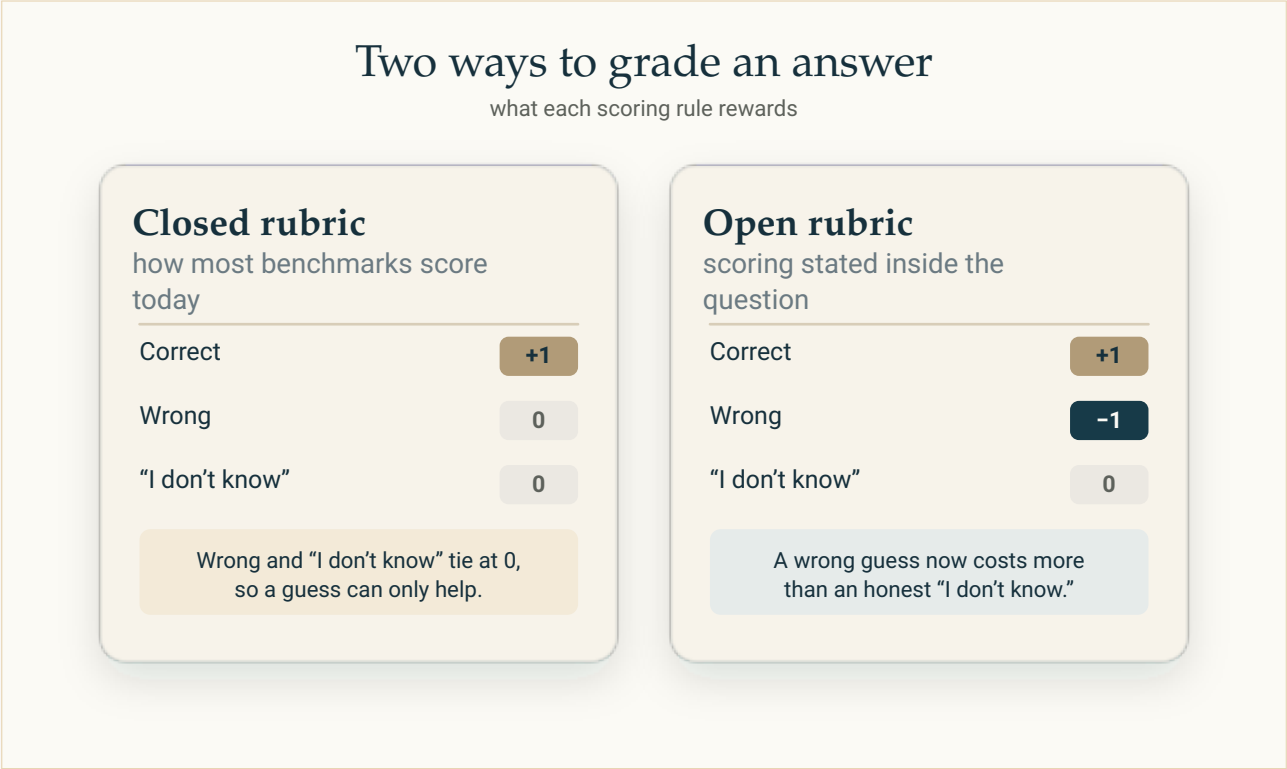
Здесь используется действительно красивая идея — старше языковых моделей и заслуживающая отдельного знакомства.

Начнем с мешка цветных шариков. Вы не знаете, сколько в нем цветов. Вы вытаскиваете 100 шариков по одному и считаете: красных 40, синих 25, зеленых 15, желтых 5, фиолетовых 3, оранжевых 2 — а затем **десять разных цветов, каждый из которых встретился ровно один раз**.

Теперь вопрос, с которым Тьюринг на самом деле столкнулся в совсем другой задаче: *какова вероятность, что следующий шарик будет цвета, которого мы вообще еще не видели?* Посчитать то, чего вы никогда не вытаскивали, невозможно — но **можно** посчитать цвета, встретившиеся ровно один раз, то есть единичные наблюдения. Трюк, известный как **оценивание Гуда–Тьюринга (Good–Turing estimation)**, состоит в том, что доля одноразовых наблюдений оценивает вероятностную массу, которая все еще скрывается в неизвестных категориях. Десять из ста вытянутых шариков принадлежали

> цветам, встретившимся лишь раз, значит, вероятность того, что следующий цвет окажется совершенно новым, составляет примерно **10 / 100 = 10%.**
> Эти одноразовые цвета — не *ошибки*. Они являются измерением вашего собственного незнания: большое число цветов, появившихся один раз, — способ, которым выборка сообщает, что в мире есть еще много того, чего вы просто не вытянули.
> **Теперь заменим цвета на дни рождения,** а мешок — на тренировочный текст модели. Предположим, среди увиденных дней рождения **каждый пятый встречается ровно один раз.** Тот же трюк: примерно пятая часть вероятностной массы приходится на дни рождения, которых модель фактически никогда не видела, а у дня рождения нет закономерности, на которую можно опереться — его нельзя *вычислить*. Поэтому дата, увиденная один раз или никогда, — монета, вес которой модель не знает, и она будет ошибаться примерно в **одном случае из пяти**. Никакая «умность» не исправит отсутствие информации: учиться было не на чем.
> В миниатюре это весь аргумент: доля единичных наблюдений измеряет, сколько мира *невозможно выучить из этих данных*, и это становится **нижней границей** ошибок. Именно поэтому модель почти никогда не промахивается со столицей: *Paris* встречается постоянно, доля единичных наблюдений для него близка к нулю, значит, информации для обучения много.

Это объясняет, откуда берутся галлюцинации. Но не объясняет, почему они *выживают* — почему после всех последующих этапов обучения, призванных сделать модель полезной и честной, она все равно блефует вместо признания неуверенности. Здесь аналогия авторов почти неловко точна. Представьте студента на экзамене, который не знает ответа. Если пустой ответ дает ноль, а догадка может дать один балл, стратегия, максимизирующая оценку, — угадывать: уверенно, конкретно, никогда не писать «я не знаю». Студенты этому учатся. Как выясняется, модели тоже — потому что мы оцениваем их так же. Авторы просмотрели бенчмарки, на которых реально соревнуется отрасль, рейтинги, под которые модели оптимизируют, и обнаружили: почти все дают «я не знаю» ровно столько же баллов, сколько неправильному ответу, — ноль. При таком правиле модель, которая всегда угадывает, обгонит идентичную модель, честно обозначающую неуверенность. В довольно буквальном смысле мы сами системой оценки подталкиваем их к этому.



Бенчмарки могут делать угадывание рациональной стратегией. При схеме оценки, которую использует большинство тестов (слева), неправильный ответ и честное «я не знаю» оба дают ноль — значит, догадка может только помочь. Если правила явно записать в самом вопросе и штрафовать за ошибку (справа), отказ от ответа при неуверенности может стать лучшим выбором. Это меняет то, что вознаграждает тест; само по себе оно не решает проблему галлюцинаций.

Original diagram — The Clean Paper CC BY 4.0

Именно эту часть стоит запомнить, потому что она противоречит привычному заголовку. Галлюцинации часто подают как неизбежную, почти мистическую границу технологии. Статья возражает обеим частям. Нижняя граница на этапе предобучения не загадка — это обычная статистическая ошибка, которую машинное обучение понимает десятилетиями. А ее сохранение не неизбежно: отчасти это стимул, который мы сами создали и можем изменить. Система, которая просто отказывается отвечать, когда не уверена, вообще не галлюцинировала бы; развернутые модели не ведут себя так потому, что наши таблицы результатов наказывают отказ.

Что сделали авторы

Статья состоит из трех частей. Первая — математический аргумент, что некоторая доля галлюцинаций статистически неизбежна уже на этапе предобучения: авторы показывают, что задача «генерировать только корректный текст» как минимум так же сложна, как бинарная классификация «это утверждение корректно или нет?». Вторая — аргумент, подкрепленный обзором десяти влиятельных бенчмарков, что обычные метрики точности вознаграждают угадывание сильнее воздержания от ответа. Третья

— предложенное исправление и иллюстративный пример: **оценивание с открытыми критериями**, где правило подсчета баллов прямо записано в вопросе (например, «правильный ответ дает 1 балл, неправильный −1, поэтому воздержитесь, если ваша уверенность ниже 50%»), чтобы модель знала, когда честность вознаграждается. Это проверяют на четырех передовых моделях — Gemini 3 Pro от Google, GPT-5 от OpenAI, Grok 4 от xAI и Claude Opus 4.5 от Anthropic — используя 4 326 фактических вопросов SimpleQA. Авторы прямо пишут, что этот пример носит лишь иллюстративный характер и «не является контролируемым сравнением моделей» — стандартные настройки, без тюнинга и нормализации стоимости.

Что они обнаружили

- **Предобучение неизбежно оставляет некоторую долю ошибок.** Частота, с которой модель выдает уверенные ложные утверждения, имеет нижнюю границу, приблизительно связанную с удвоенной частотой ошибки лучшего классификатора «корректно ли это утверждение?», построенного из нее. Для фактов без закономерности, которую можно выучить, эта граница как минимум равна доле единичных наблюдений — доле фактов, встречающихся в обучении ровно один раз. Некоторые галлюцинации неизбежны даже при совершенно чистых данных.
- **Оценивание конкретно вознаграждает угадывание.** При обычной схеме «правильно/непра оптимальная стратегия — никогда не воздерживаться от ответа, а обзор авторов показывает, что подавляющее большинство популярных бенчмарков засчитывает «я не знаю» просто как ошибку. Яркий пример из их собственных моделей: на SimpleQA сырая точность слегка *предпочитает* OpenAI o4-mini, которая отвечает почти на все и ошибается более чем в трех четвертях случаев, GPT-5-mini, которая делает намного меньше ошибок, потому что воздерживается, когда не уверена. Более безрассудная модель выглядит лучше в таблице.
- **Открытые критерии оценивания меняют стимул — в их иллюстративном примере.** Авторы тестируют простой способ уменьшения галлюцинаций: попросить модель ответить дважды и воздержаться, если два ответа не совпадают. При стандартной метрике точности метод уменьшает число ошибок, *но одновременно снижает сам показатель точности*, поэтому метрика фактически наказывает его использование. При открытых критериях оценивания тот же метод оказывается лучше для всех четырех моделей в широком диапазоне штрафов; а GPT-5-mini, которую сырая метрика точности наказывала за отказ от ответа при неуверенности, обгоняет o4-mini, когда схема баллов сформулирована открыто (n = 4 326 вопросов на модель).

Что это, вероятно, означает

Снижение галлюцинаций в основном не требует придумывать еще больше специальных тестов на галлюцинации. Нужно изменить то, как главные бенчмарки оценивают неуверенность, чтобы признание «я не знаю» перестало наказываться. Пока таблица результатов не изменится, уменьшение галлюцинаций по-прежнему будет стоить моделям баллов по метрике точности и потому оставаться невыгодным — именно поэтому авторы называют проблему «социотехнической»: нужна и лучшая метрика, и готовность влиятельных рейтингов ее принять.

Чего эта работа *не* доказывает

- Она **не** показывает, что открытые критерии оценивания устраняют галлюцинации в реальном использовании. Поддерживающий эксперимент — маленький и намеренно **неконтролируемый** иллюстративный пример: четыре модели со стандартными настройками, один выбранный способ уменьшения галлюцинаций, один тест фактических вопросов. Его цель — показать *переворот стимула*, а не ранжировать модели или доказать общую эффективность.
- Она **не** утверждает, что оценивание — *единственная* причина. Ошибки в тренировочных данных, действительно сложные задачи и незнакомые запросы остаются отдельными источниками.
- Она **не** поддерживает популярное утверждение, что галлюцинации неизбежны. Авторы утверждают противоположное: система, которая отвечает только на проверяемые вопросы и иначе говорит «я не знаю», никогда не галлюцинировала бы.
- Она **не** убирает нижнюю границу, возникающую на этапе предобучения — она ее объясняет и ограничивает, а сама граница касается уверенных фактических ошибок, не всего поведения модели.
- Она **не** показывает, что открытых критериев оценивания *достаточно* сами по себе. Они меняют то, что вознаграждает оценивание; они не заменяют поиск по внешним источникам, использование инструментов или лучше откалиброванные модели.

Насколько убедительны доказательства?

- Ядро работы — **математика**: формальные нижние границы, а не измерения. Как теоретический аргумент она последовательна в рамках собственных предположений.
- Она опирается на **намеренно упрощенные модели** проблемы. Сами авторы отмечают «ложную трихотомию» — разделение каждого ответа только на правильный, неправильный или «я не знаю» — и идеализированную среду «произвольных

фактов», использованную для самой чистой границы.
- Обзор бенчмарков — **маленькая, вручную отобранная выборка**: десять влиятельных оценок, а не полный аудит.
- Иллюстративный эксперимент **реален, но ограничен**: четыре передовые модели, один способ снижения галлюцинаций, только SimpleQA, стандартные настройки, и прямо сказано «не контролируемое сравнение». Это проверка концепции для аргумента о стимулах, а не новый рейтинг моделей.
- Стоит назвать и позицию авторов: трое из четырех работают или работали в **OpenAI**, а статья аргументирует, как отрасли следует изменить оценивание моделей. Это хорошо аргументированная позиция заинтересованной стороны, а не нейтральный внешний обзор — ее следует взвешивать, а не отбрасывать. (К чести статьи, она направляет критику на собственные модели o4-mini и GPT-5-mini столь же охотно, как и на другие.)

Почему это важно

Работа переосмысливает чрезвычайно раскрученную проблему. «Галлюцинацию» часто продают либо как жуткий дефект, либо как неподвижную стену; эта статья делает ее более обычной и частично самосозданной — статистическая нижняя граница, которую мы можем понять, плюс стимул, который мы сами выбрали. Более широкий урок тише и полезнее: дальнейший прогресс в надежности может зависеть не меньше от *того, что мы измеряем*, чем от того, что мы строим.

Краткий итог

Уверенные ложные ответы языковых моделей имеют два источника. Первый статистический: когда факт не имеет закономерности, которую можно выучить, модель, обученная имитировать язык, иногда будет ошибаться, и эту нижнюю границу можно оценить — например, по доле фактов, встречающихся в обучении лишь один раз. Второй источник — стимулы: почти каждый бенчмарк, по которому ранжируют модели, оценивает «я не знаю» так же, как неправильный ответ, поэтому угадывать всегда выгоднее — вплоть до ситуации, когда модель, ошибающаяся в трех четвертях случаев, может обогнать более честную модель, которая воздерживается от ответа. Предложение авторов — не еще один тест на галлюцинации, а **открытые критерии оценивания**: записывать правило оценивания прямо в вопросе. В иллюстративном эксперименте на четырех передовых моделях это переворачивает стимул, и метод уменьшения галлюцинаций вознаграждается, а не наказывается. Это теоретическая работа плюс обзор с небольшим, явно неконтролируемым экспериментом, рецензированная в *Nature*. Исправление перспективно, но еще не показано в большом масштабе; сами галлюцинации здесь трактуются как не мистические и не строго неизбежные.

Проверка без прикрас

Что показывает статья: Математическую нижнюю границу, при которой некоторые галлюцинации возникают уже на этапе предобучения — для фактов без закономерностей как минимум на уровне доли единичных наблюдений; обзор, согласно которому большинство ведущих бенчмарков не дает никакого бонуса за «я не знаю»; и иллюстративный эксперимент на четырех моделях, где явное формулирование правил оценки в запросе — открытые критерии оценивания — делает метод уменьшения галлюцинаций выгодным там, где обычная метрика точности его наказывала.

Что правдоподобно, но не доказано: Что внедрение открытых критериев оценивания в главные бенчмарки существенно уменьшит галлюцинации в развернутых моделях. Поддерживающий эксперимент невелик и прямо назван неконтролируемым.

Чего она не показывает: Что галлюцинации неизбежны — работа утверждает противоположное; что оценивание является единственной причиной; что галлюцинации можно полностью устранить; что этот иллюстративный эксперимент ранжирует четыре модели между собой.

Главные ограничения: Намеренно упрощенная модель «правильно/неправильно/я не знаю», которую сами авторы называют «ложной трихотомией»; небольшой вручную отобранный обзор бенчмарков — десять оценок; неконтролируемый иллюстративный эксперимент — четыре модели, один способ снижения галлюцинаций, один тест, стандартные настройки; и аргумент, в значительной мере подготовленный исследователями OpenAI, о том, как отрасли следует оценивать модели.

Какой уверенности заслуживает вывод для обычного читателя? Высокой, что галлюцинации не являются ни мистическими, ни строго неизбежными и что главные бенчмарки сейчас вознаграждают угадывание. Умеренной, что предложенное исправление поможет: теперь есть настоящая проверка концепции, но еще нет демонстрации широкой эффективности в большом масштабе.

Источники

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