The Clean Paper

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

语言模型为何产生幻觉——而我们的评分方式为何让它延续

我们所说的“幻觉”，即模型自信地输出虚假内容，并不是什么神秘故障：其中一部分是训练在统计上的副产品，而它们之所以持续存在，是因为主流基准会奖励自信猜测，却不会奖励诚实地说“我不知道”。一项针对四个前沿模型的案例研究表明，把评分规则写进提示（“开放式评分准则”）可以扭转这种激励。

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, 1047–1050 (2026) · DOI: 10.1038/s41586-026-10549-w

模型为何猜测，以及我们为何把它们训练成这样

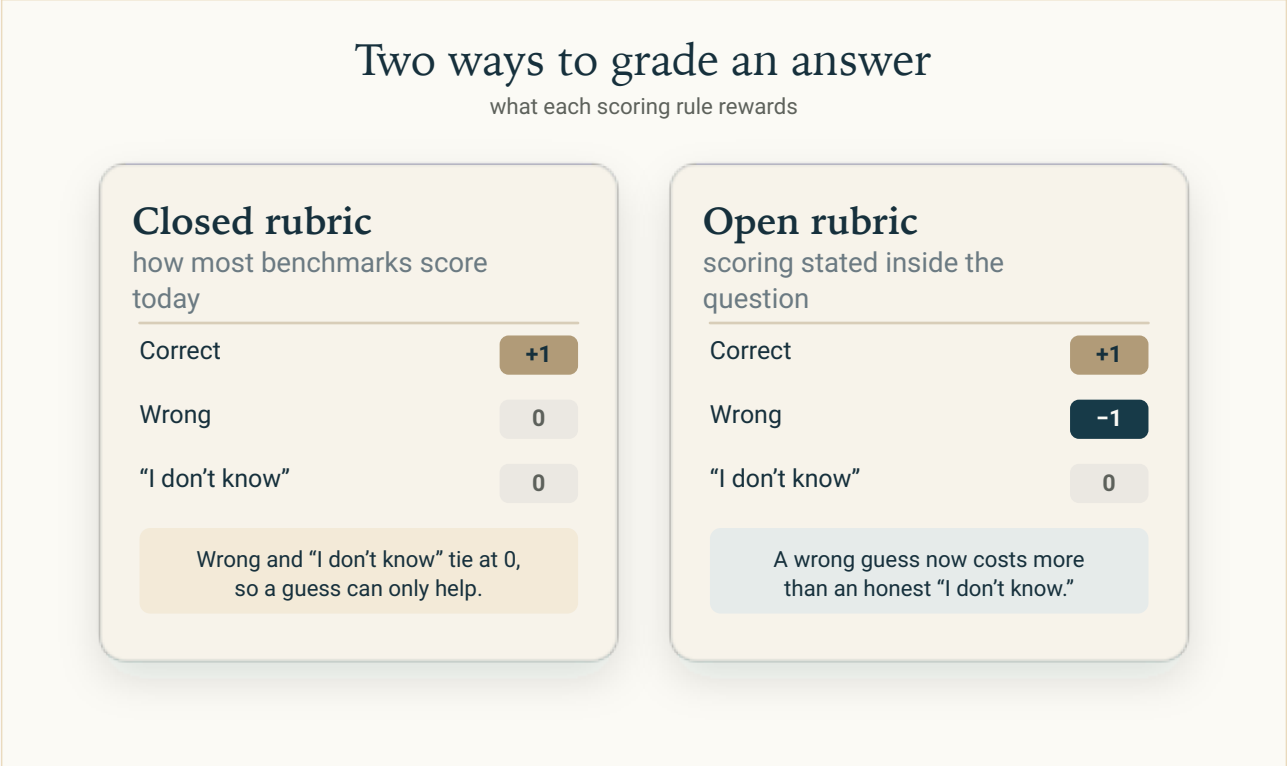
如果你问一个大型语言模型某个陌生人的生日，它可能会以仿佛照着卡片念出的笃定口吻回答“3 月 7 日”——但连续问三次，得到三个不同的日期，而且全都错。作者给出的正是这类例子：让领先模型回答一个朴素的事实问题——某人的生日，或某个冷僻缩写代表什么——每个模型都自信地编出不同答案，却没有一个正确。业界把这种现象称为幻觉，听起来像是感知出了故障。论文的第一步，就是为它祛除神秘色彩。

先看模型是怎样建成的。在第一个也是规模最大的训练阶段，模型通过阅读海量文本，实际上是在学习流畅语言长什么样。现在考虑一个背后没有规律的事实——某个特定人物的生日。如果那个日期在训练文本里只出现过一次，甚至从未出现，模式学习器就无从抓取规律：从模型的角度看，答案是任意的。作者借用一个旧思想（艾伦·图灵在另一个问题上提出的思想）把这一点说得更精确：如果五个生日里有一个在数据中只出现一次，那么模型至少答错五分之一，应当是意料之中的事——不是因为模型坏了，而是因为原本就没有可供学习的东西。（同理，模型几乎不会答错一个国家的首都：这些事实会反复出现。）作者谨慎论证，区分真实陈述与貌似可信的虚假陈述本身就是难题，而生成全部为真的陈述至少同样困难。错误下限从一开始就内置其中。

> **核心思想：怎样计算那些你还没见过的东西**
>
> 这建立在一个真正巧妙的思想上——它比语言模型古老得多，值得认真认识。
> **从一袋彩球开始。**你不知道袋里有多少种颜色。你逐个抽出 100 个并计数：红色 40 个、蓝色 25 个、绿色 15 个、黄色 5 个、紫色 3 个、橙色 2 个——然后还有**十种不同的颜色，每种都只出现一次**。
> 图灵当时面对的是另一个问题：下一颗球是一种你从未见过的颜色，概率是多少？你无法计算从未抽到过的颜色——但你**可以**计算只见过一次的颜色，即“单例”。这种称为**古德–图灵估计**的技巧认为：单例在全部抽样中所占的比例，可以估计仍隐藏在未见颜色里的概率质量。100 次抽取中有 10 次属于只出现一次的颜色，因此下一颗球是全新颜色的概率约为 **10 / 100 = 10%**。
> 那些只见过一次的颜色不是错误。它们是在衡量你自己的无知：许多颜色只出现一次，是样本在告诉你，世界上还有更多颜色，只是你尚未抽到。
> **现在把颜色换成生日，**把袋子换成模型的训练文本。假设在模型见过的生日中，**五个里有一个只出现一次。**同样的技巧表明：大约五分之一的概率质量落在模型实际上从未见过的生日上——而生日没有可供兜底的规律（你无法推算出某人的生日）。因此，面对一个只见过一次或从未见过的日期，模型就像面对一枚无法判断偏向的硬币；大约**五个就会错一个**。再聪明也无法解决这一点：那里本来就没有东西可学。
> 这就是整个论证的缩影：单例率衡量这个世界有多少内容无法从这批数据中学到，并由此形成错误的**下限**。这也解释了模型为何几乎不会答错首都——巴黎反复出现，它的单例率接近于零，因此有充足的信息可学。

这解释了幻觉从何而来，却没有解释它们为何会留存下来——为什么模型经过后来所有旨在让它更有帮助、更诚实的训练之后，仍会虚张声势，而不是承认不确定。论文在这里用了一个几乎令人不适地贴切的类比。设想一个学生在考试中不知道答案。如果留空得零分，而猜测有可能得一分，那么追求最高分的做法就是猜——要自信、要具体，绝不能说“我不确定”。学生会学会这一点。事实证明，模型也会——因为我们的评分方式完全相同。作者检查了这个领域真正竞争的基准，也就是模型经过调优要攀登的排行榜，发现几乎所有基准都把“我不知道”和错误答案同样计为零分。在这条规则下，总是

猜测的模型会击败另一个除了会诚实标示不确定性之外完全相同的模型。从相当字面的意义上说，是我们的评分把它们推向了幻觉。



基准测试可能让猜测成为理性选择。在多数基准采用的评分方式下（左），错误答案和诚实的 “我不知道” 都得零分——因此猜测只可能有利。若在问题中写明规则，并对错误答案扣分（右），那么在不确定时放弃作答就可能成为更优选择。这改变了测试所奖励的行为；但它本身并不能解决幻觉。

Original diagram —The Clean Paper CC BY 4.0

这部分最值得记住，因为它与常见标题背道而驰。幻觉经常被描述成技术不可避免、近乎神秘的极限。论文在这两点上都提出异议。预训练中的错误下限并不神秘——它只是普通的统计误差，是机器学习几十年来一直理解的那种问题。幻觉的持续存在也并非不可避免——它在一定程度上源于我们设计、也能够改变的激励。一个在不确定时干脆拒绝回答的系统根本不会产生幻觉；已部署模型不这样做，是因为我们的记分牌会惩罚拒答。

作者做了什么

论文分为三个部分。第一，作者用数学论证说明，预训练期间某些幻觉在统计上不可避免：他们证明，“只生成有效文本” 至少与 “这条陈述是否有效” 的二元分类问题一样困难。第二，作者调查十项有影响力的基准，并据此论证，主流的准确率式指标奖励猜测而不是弃答。第三，作者提出修正方案并用案例研究检验它：采用**开放式评分准则**，把评分规则写在问题本身之中（例如，“答对得 1 分，答错得 −1 分，因此若把握低于 50% 就应弃答”），让模型知道何时诚实会得到奖励。他们在谷歌 Gemini 3 Pro、OpenAI GPT-5、xAI Grok 4 和 Anthropic Claude Opus 4.5 四个前沿模型上测试这一做法，使用SimpleQA的 4,326 个事实问题。作者明确表示，这项案例研究只作说明，“不是跨模型的受控评估”

（使用默认设置，没有调优，也没有成本归一化）。

他们发现了什么

- **预训练必然造成一些错误**。模型生成自信虚假陈述的比率存在一个下限，大致是由该模型构建的最佳“这条陈述是否有效”分类器错误率的两倍。对于没有可学习规律的事实，这个下限至少等于单例率——即训练中只出现一次的事实所占比例。即使数据完全干净，一些幻觉仍不可避免。
- **评分确实奖励猜测**。在普通的对错评分下，从不弃答是最优策略；作者的调查发现，绝大多数常用基准都把“我不知道”直接算作错误。他们自己的一个鲜明例子是：在 SimpleQA 测试上，原始准确率略微偏爱 OpenAI 的 o4-mini——它几乎每题都答，却有四分之三以上答错——而不是 GPT-5-mini；后者会在不确定时弃答，因此犯错少得多。更冒进的模型反而在记分牌上看起来更好。
- **开放式评分准则扭转了激励（在这项案例研究中）**。作者测试了一种简单的幻觉缓解方法：让模型回答两次，如果两个答案不一致就弃答。在标准准确率下，这种方法减少了错误，但也降低了准确率——因此指标会阻碍人们采用它。在开放式评分准则下，同一种方法在一系列错误惩罚设置中，对四个模型都更占优势；原始准确率曾因 GPT-5-mini 在不确定时弃答而惩罚它，但规则公开后，它也领先于 o4-mini（每个模型 n = 4,326 个问题）。

这可能意味着什么

减少幻觉，主要并不是发明更多针对幻觉的专门测试，而是改变主流基准给不确定性打分的方式，让承认“我不知道”不再受到惩罚。只要记分牌不变，减少幻觉就会继续让模型损失准确率分数，从而继续受到阻碍——这正是作者把问题称为“社会技术”问题的原因：一部分需要更好的指标，一部分需要让有影响力的排行榜采纳它。

这项研究没有证明什么

- 它**没有**表明开放式评分准则能修复真实环境中的幻觉。支撑这一主张的实验是一项小型、刻意保持**非受控**的案例研究——四个模型、默认设置、一种选定的缓解方法、一项事实问答测试——目的是展示激励如何翻转，而不是给模型排名或证明普遍有效。
- 它**没有**声称评分是唯一原因。训练数据中的错误、真正困难的问题和陌生提示仍是彼此独立的来源。
- 它**不支持**“幻觉不可避免”这句流行说法。作者的论点恰好相反：一个只回答可核查问题、否则就说“我不知道”的系统永远不会产生幻觉。
- 它**没有**让预训练的错误下限消失——它解释并界定了这个下限，而且该界限针对的是自信的事实错误，不是模型的全部行为。
- 它**没有**证明开放式评分准则单独使用就已足够。它们改变评估所奖励的行为，却不能替代检索、工具使用或校准更好的模型。

证据有多强?

- 核心证据是**数学——**正式的下界，而不是测量结果。作为理论论证，它在自身假设内成立。
- 它建立在**刻意简化的问题模型**上；作者自己指出，把每个回答归为正确、错误或“我不知道”是一种“虚假的三分法”，而最清晰的下界也采用了理想化的“任意事实”设定。
- 基准调查是一个**小型、经筛选的样本——**十项有影响力的评估，不是穷尽式审计。
- 案例研究**真实但有限**：四个前沿模型、一种缓解方法、仅 SimpleQA、默认设置，并明确注明“不是受控评估”。它是激励论点的概念验证，不是基准测试结果。
- 还应说明其立场：四位作者中有三位正在或曾经受雇于 **OpenAI**，而论文主张该领域应改变模型评估方式。这是一项由利益相关方提出、论证充分的立场，而不是中立的外部审查——应当权衡，而非因此否定。（值得肯定的是，论文对自家模型 o4-mini 和 GPT-5-mini 的批评，与对其他模型同样直接。）

为什么重要

它重新界定了一个被严重炒作的问题。“幻觉”通常要么被包装成诡异的缺陷，要么被说成无法跨越的高墙；这篇论文则把它还原为普通现象，而且部分是我们自己造成的——一个可以理解的统计下限，叠加在我们亲手选择的激励之上。更广泛的教训更平实，也更有用：可靠性的进一步提升，可能既取决于我们衡量什么，也取决于我们建造什么。

清晰摘要

语言模型自信地给出错误答案，来自两个方面。第一是统计因素：当一个事实没有可学的规律时，经过语言模仿训练的模型有时会答错，而这个下限可以估计（例如依据训练中有多少事实只出现一次）。第二是激励因素：几乎所有用于模型排名的基准，都把“我不知道”和错误答案打成同样的分数，所以猜测总会占优——甚至一个有四分之三时间答错的模型，也可能胜过一个会弃答、因而更诚实的模型。作者提出的不是另一种幻觉测试，而是“开放式评分准则”：在问题里写明评分方式。在四个前沿模型的案例研究中，这扭转了激励，使一种减少幻觉的方法获得奖励，而不再受到惩罚。这是一篇结合理论与调查、并附带一项明确注明非受控小型实验的论文，已在 *Nature* 接受同行评审；所提方案前景可期，但尚未证明能大规模奏效，而作者认为幻觉既不神秘，也并非严格不可避免。

不绕弯核查

论文证明了什么：预训练期间某些幻觉不可避免的数学下界（对于没有规律的事实，至少是“单例率”）；一项调查发现，多数领先基准不给“我不知道”任何分数；以及一项四模型案例研究，其中把评分规则写进提示（“开放式评分准则”），会让减少幻觉的方法胜出，而普通准确率原本会惩罚它。

什么合理但尚未证明：如果主流基准采用开放式评分准则，已部署模型的幻觉会显著减少。支撑实验

规模很小，而且明确属于非受控研究。

它没有证明什么：幻觉不可避免（它主张相反）；评分是唯一原因；幻觉可以被彻底消除；这项案例研究能给四个模型排出高下。

主要局限：刻意简化的正确/错误/“我不知道”模型（作者称之为“虚假的三分法”）；小型、经筛选的基准调查（十项评估）；非受控案例研究（四个模型、一种缓解方法、一项测试、默认设置）；以及一项由 OpenAI 主导、讨论该领域应如何评估模型的论证。

普通读者应该有多大信心？可以高度确信幻觉既不神秘，也并非严格不可避免，而且主流基准目前确实奖励猜测。对于所提修正方案能否奏效，应持中等信心——它已有真实的概念验证，但尚未证明可以广泛、大规模地发挥作用。

来源

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