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 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

为什么模型会猜，以及我们为什么教会了它们这样做

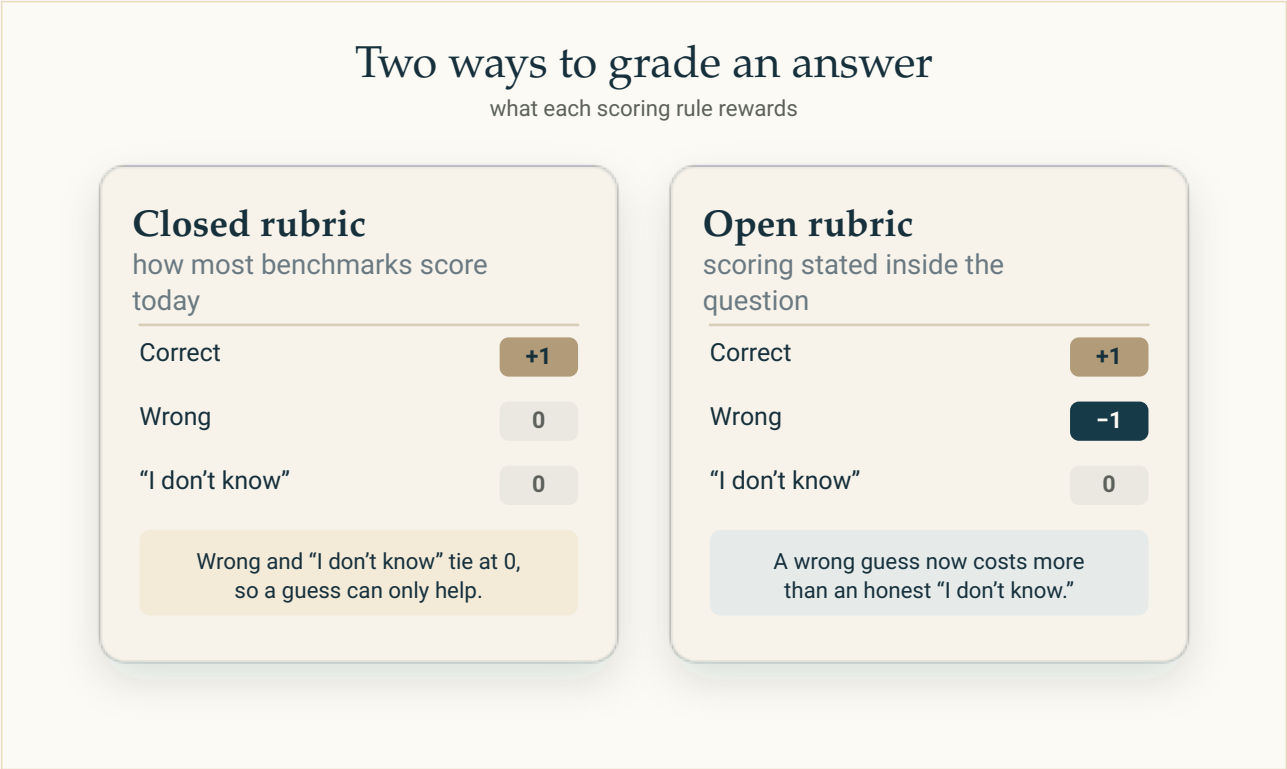
向一个大型语言模型询问陌生人的生日，它可能会用一种仿佛正在从卡片上读出答案的笃定语气回答“3 月 7 日”——但它也可能答错，甚至连续三次给出三个不同的日期。作者给出了正是这种类型的例子：向领先模型提出一个简单的事实问题——某人的生日，或某个冷门缩写代表什么——每个模型都自信地编造出一个不同的答案，而且没有一个正确。这个行业把它称为幻觉，听起来仿佛是感知出了问题。论文的第一步，就是去掉其中的神秘色彩。

先从模型的构建方式说起。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的语言是什么样的。现在考虑一个背后没有任何模式的事实——某个特定人的生日。如果这个日期在训练文本中只出现过一次，或者从未出现过，那么作为模式学习者的模型就没有任何东西可以抓住：从模型的角度看，这个答案是任意的。作者借用了一个更早的观点（艾伦·图灵针对另一个问题提出的观点），并将其精确化：如果五分之一的生日在数据中只出现一次，那么模型至少应当答错其中五分之一——这不是因为它出了故障，而是因为原本就没有任何东西可供学习。（按照同样的逻辑，模型几乎从不会答错一个国家的首都：这些答案会反复出现。）他们谨慎地论证，辨别一个真实陈述和一个貌似合理的错误陈述本身就是难题，而只生成真实陈述至少同样困难。错误的下限已经写进系统之中。

> **背后的想法：如何计算你尚未见过的东西**
>
> 这建立在一个真正巧妙的想法之上——它早于语言模型，值得认真了解。
> **先从一袋彩色球开始。**你不知道袋子里有多少种颜色。你逐个抽取 100 个球并记录：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**十种各自恰好出现一次的不同颜色**。
> 现在来看图灵实际面对的问题，虽然那是一个完全不同的问题：下一个球是你完全没见过的颜色的概率是多少？你无法数出从未抽到过的东西——但你**可以**数出那些恰好见过一次的颜色，也就是“单例”。这个叫作 **Good–Turing 估计**的技巧是：抽取结果中单例所占的比例，可以用来估计尚未见过的颜色仍然隐藏着的概率。100 次抽取中有 10 次是只出现过一次的颜色，所以，下一个球是全新颜色的概率约为 **10 / 100 = 10%**。
> 那些只见过一次的颜色并不是错误。它们测量的是你自己的无知：许多颜色只出现一次，是样本在告诉你，世界中还有更多东西，只是你还没有抽到。
> **现在把颜色换成生日，**把袋子换成模型的训练文本。假设在它见过的生日中，**五分之一恰好只出现一次。**同样的技巧表明：大约五分之一的概率属于模型实际上从未见过的生日——而生日没有可以依赖的模式（你无法推算出某人的生日）。所以，一个只见过一次、或从未见过的日期，就像模型无法赋予权重的一枚硬币；对于其中大约**五分之**一的日期，它会答错。再聪明也解决不了这个问题：原本就没有任何东西可供学习。
> 这就是整个论证的缩影：单例率衡量了世界中有多少内容无法从这些数据中学到，而它会成为错误的一个**下限**。这也解释了模型几乎不会漏掉一个首都——*Paris* 会不断出现，它的单例率接近于零，因此有大量东西可供学习。

这解释了幻觉从何而来，却没有解释它们为什么会存活下来——为什么模型在经历了所有旨在让它们更有帮助、更诚实的后续训练后，仍然会虚张声势，而不是承认自己不确定。论文在这里使用的类比几乎贴切得令人不安。想象一名学生在考试中遇到不会的问题。如果空白答案得零分，而猜测可能得一分，那么最大化成绩的做法就是猜——要自信、具体，绝不说“我不确定”。学生会学会这一点。事

实证明，模型也会——因为我们用同样的方式给它们评分。作者梳理了这个领域实际竞争的基准测试，也就是模型被调校来攀登的排行榜，发现其中几乎所有测试都把“我不知道”和错误答案打出完全相同的分数：零分。在这条规则下，一个总是猜测的模型会胜过一个其他方面完全相同、但诚实标记自身不确定性的模型。从相当字面意义上说，是我们用评分把它们训练成这样的。



基准测试可能会让猜测变得理性。在大多数基准测试采用的评分方式下（左），错误答案和诚实的“我不知道”都得零分——因此猜测只有可能带来帮助。在问题中写明规则，并对答错进行惩罚（右），那么不确定时弃答就可能成为更好的选择。这会改变测试所奖励的行为；但它本身并不能解决幻觉。

这一点值得保留，因为它与通常的头条说法相悖。人们常把幻觉描述为这项技术不可避免、近乎神秘的局限。论文在这两个方面都提出了异议。预训练下限并不神秘——它就是普通的统计误差，是机器学习几十年来已经理解的那类误差。而幻觉的持续存在也不是不可避免的——它部分源于我们建立、并且可以改变的一种激励。一个在不确定时直接拒绝回答的系统根本不会产生幻觉；已部署的模型之所以不这样做，是因为我们的排行榜会惩罚拒答。

作者做了什么

这篇论文分为三部分。第一部分是一个数学论证：通过证明“只生成有效文本”至少和二元的“这条陈述是否有效？”分类问题一样难，说明预训练期间的一些幻觉在统计上是被迫出现的。第二部分是一个论证，并由对十个有影响力的基准测试的调查提供支持：主流的准确率类指标奖励猜测，而不是弃答。第三部分提出一个修复方案并进行案例研究：**open-rubric** 评估，在这种评估中，评分规则直接写在问题内部（例如，“正确答案得 1 分，错误答案得 −1 分，因此如果你的确信度低于 50%，就弃答”），这

样模型就能判断诚实是否会得到奖励。他们在四个前沿模型上进行了尝试——Google 的 Gemini 3 Pro、OpenAI 的 GPT-5、xAI 的 Grok 4 和 Anthropic 的 Claude Opus 4.5——使用了SimpleQA的 4,326 个事实问题。他们明确说明，这项案例研究只是示例性的，“不是跨模型的受控评估”（使用默认设置、未调优、未进行成本归一化）。

他们发现了什么

- **预训练会迫使模型犯下一些错误**。模型自信地输出错误陈述的比率，下限大致是由它构建出的最佳 “这条陈述是否有效？” 分类器错误率的两倍。对于没有可学习模式的事实，这个下限至少是单例率——训练中恰好出现一次的事实所占的比例。即使数据完全无噪声，一些幻觉也不可避免。
- **评分确实会奖励猜测**。在通常的正确/错误评分下，从不弃答是最优策略；作者的调查发现，绝大多数热门基准测试都会把 “我不知道” 直接当作错误。作者自己的结果提供了一个鲜明例子：在 SimpleQA 测试中，原始准确率稍微偏向 OpenAI 的 o4-mini——它几乎回答所有问题，而且超过四分之三的答案是错的——而不是 GPT-5-mini；后者因为在不确定时弃答，犯错少得多。更鲁莽的模型在排行榜上看起来更好。
- **开放评分规则会扭转激励（在他们的案例研究中）**。他们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，这种缓解方法会减少错误，但也会降低准确率——因此该指标会阻止采用它。在开放评分规则下，同一种缓解方法在一系列惩罚设置下对四个模型都取得更好结果；而 GPT-5-mini——原始准确率会因为它在不确定时弃答而惩罚它——在公开说明评分后超过了 o4-mini（每个模型的题目数 n = 4,326）。

这可能意味着什么

减少幻觉主要不是发明更多针对幻觉的测试，而是改变主流基准测试衡量不确定性的方式，让承认 “我不知道” 不再受到惩罚。在排行榜改变之前，减少幻觉会继续让模型损失准确率分数，也就会继续受到阻碍——这就是作者把问题称为 “社会技术问题” 的原因：一部分是改进指标，另一部分是推动有影响力的排行榜采用它。

这并不能证明什么

- 它并**没有**表明开放评分规则能修复现实环境中的幻觉。作为支持证据的实验规模很小，而且是刻意设置的**非受控**案例研究——四个使用默认设置的模型、一种选定的缓解方法、一次事实问答测试——目的是展示激励的翻转，而不是给模型排名或证明普遍有效性。
- 它并**没有**声称评分是唯一的原因。训练数据中的错误、真正困难的问题以及陌生的提示，仍然是独立的来源。
- 它并**不支持** “幻觉不可避免” 这一流行说法。作者的论点恰恰相反：一个只回答可核查问题、其他时候都说 “我不知道” 的系统永远不会产生幻觉。
- 它并**没有**让预训练下限消失——它解释并界定了这个下限，而且这个界限针对的是自信的事实错

误，并非模型的全部行为。
- 它并**没有**表明开放评分规则单独就足够。它们改变了评估所奖励的行为；它们不能替代检索、工具使用或校准得更好的模型。

证据有多强?

- 核心是**数学**——是形式化的下界，而不是测量结果。作为理论论证，它在自身的前提下是稳固的。
- 它建立在对问题的**刻意简化的模型**之上；作者自己也指出了把每个回答都视为正确、错误或“我不知道”的“错误三分法”，以及为得到最清晰的界限而采用的理想化“任意事实”设定。
- 基准测试调查是一个**规模较小、经过筛选的样本**——十个有影响力的评估，而不是一次穷尽式审计。
- 案例研究**真实但有限**：四个前沿模型、一种缓解方法、仅使用 SimpleQA、默认设置，并明确“不是受控评估”。它是对激励论点的概念验证，不是一次基准测试结果。
- 值得说明其观察角度：四位作者中有三位现在或曾经受雇于 **OpenAI**，而论文主张这个领域应改变评估模型的方式。这是一个来自利益相关方、论证充分的立场，而不是中立的外部评审——需要加以权衡，而不是直接否定。（值得肯定的是，论文对自身的模型 o4-mini 和 GPT-5-mini 的批评，与对其他模型的批评同样直接。）

为什么这很重要

它重新界定了一个被大肆宣传的问题。“幻觉”往往被包装成一种诡异的缺陷，或一堵不可移动的墙；这篇论文则让它变得平常，而且部分是我们自己造成的——一个我们确实能够理解的统计下限，叠加在一种由我们选择的激励之上。更广泛、也更有用的启示是：可靠性要进一步提升，可能既取决于我们衡量什么，也取决于我们构建什么。

简明总结

语言模型自信的错误答案来自两个方面。第一个是统计方面：当一个事实没有可供学习的模式时，一个接受语言模仿训练的模型有时会答错，而这个下限可以被估计（例如，根据有多少事实在训练中只出现一次）。第二个是激励方面：几乎所有用于给模型排名的基准测试，都把“我不知道”和错误答案打出相同分数，因此猜测总会获胜——以至于一个四分之三的答案都错的模型，可能超过一个更诚实、会弃答的模型。作者提出的不是另一种幻觉测试，而是“open rubrics”：把评分规则写在问题内部。在对四个前沿模型进行的案例研究中，这会扭转激励，使一种减少幻觉的方法得到奖励，而不是受到惩罚。这是一篇“理论＋调查”论文，包含一个规模较小、明确为非受控的实验，并已在 *Nature* 上经过同行评审；这个修复方案很有希望，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。

不绕弯检查

论文展示了什么：一个数学下界，表明预训练期间的一些幻觉是被迫出现的（对于没有模式的事实，至少达到“单例率”）；一项调查发现，大多数领先基准测试不给“我不知道”任何分数；以及一项四模型案例研究，在其中，把评分规则写进提示词（“开放评分规则”）会让一种减少幻觉的方法胜出，而普通准确率原本会惩罚它。

合理但尚未证明的内容：如果把开放评分规则纳入主流基准测试，它们会切实减少已部署模型中的幻觉。作为支持证据的实验规模很小，而且明确是非受控的。

它没有展示什么：幻觉不可避免（它论证的是相反观点）；评分是唯一原因；幻觉可以被彻底消除；该案例研究并非用于对四个模型彼此进行排名。

主要局限：刻意简化的正确/错误/“我不知道”模型（作者称之为“错误三分法”）；规模较小、经过筛选的基准测试调查（十项评估）；非受控案例研究（四个模型、一种缓解方法、一次测试、默认设置）；以及一篇由 OpenAI 主导、讨论这个领域应如何评估模型的论证。

普通读者应有多大信心？对于“幻觉既不神秘，也并非严格不可避免”，以及“主流基准测试目前会奖励猜测”，可以有较高信心。对于所提修复方案确实有帮助，可以有中等信心——它现在有了真实的概念验证，但还没有证明能广泛、大规模地奏效。

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