The Clean Paper

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

語言模型為何產生幻覺——而我們的評分方式為何讓它延續

我們所說的「幻覺」，即模型自信地輸出虛假內容，並不是什麼神秘故障：其中一部分是訓練在統計上的副產品，而它們之所以持續存在，是因為主流基準會獎勵自信猜測，卻不會獎勵誠實地說「我不知道」。一項針對四個尖端模型的案例研究表明，把評分規則寫進提示（「開放式評分準則」）可以扭轉這種誘因。

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

模型為何猜測，以及我們為何把它們訓練成這樣

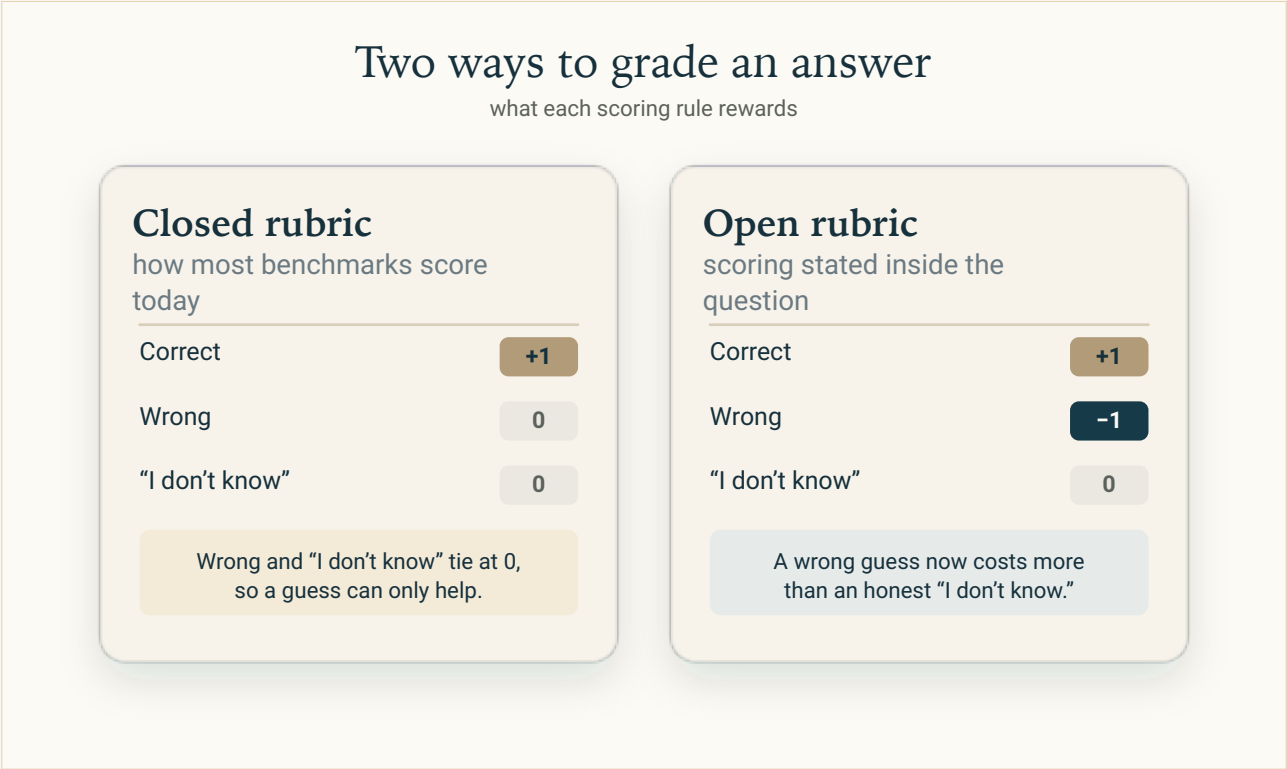
如果你問一個大型語言模型某個陌生人的生日，它可能會以彷彿照著卡片念出的篤定口吻回答「3 月 7 日」——但連續問三次，得到三個不同的日期，而且全都錯。作者給出的正是這類例子：讓頂尖模型回答一個樸素的事實問題——某人的生日，或某個冷僻縮寫代表什麼——每個模型都自信地編出不同答案，卻沒有一個正確。業界把這種現象稱為幻覺，聽起來像是感知出了故障。論文的第一步，就是為它祛除神秘色彩。

先看模型是怎樣建成的。在第一個也是規模最大的訓練階段，模型透過閱讀海量文字，實際上是在學習流暢語言長什麼樣。現在考慮一個背後沒有規律的事實——某個特定人物的生日。如果那個日期在訓練文字中只出現過一次，甚至從未出現，模式學習器就無從抓取規律：從模型的角度看，答案是任意的。作者借用一個舊思想（艾倫·圖靈在另一個問題上提出的思想）把這一點說得更精確：如果五個生日中有一個在資料中只出現一次，那麼模型至少答錯五分之一，應當是意料之中的事——不是因為模型壞了，而是因為原本就沒有可供學習的東西。（同理，模型幾乎不會答錯一個國家的首都：這些事實會反覆出現。）作者謹慎論證，區分真實陳述與貌似可信的虛假陳述本身就是難題，而生成全部為真的陳述至少同樣困難。錯誤下限從一開始就內置其中。

> **核心思想：怎樣計算那些你還沒見過的東西**
>
> 這建立在一個真正巧妙的思想上——它比語言模型古老得多，值得認真認識。
> **從一袋彩球開始**。你不知道袋裡有多少種顏色。你逐個抽出 100 個並計數：紅色 40 個、藍色 25 個、綠色 15 個、黃色 5 個、紫色 3 個、橙色 2 個——然後還有**十種不同的顏色，每種都只出現一次**。
> 圖靈當時面對的是另一個問題：下一顆球是一種你從未見過的顏色，機率是多少？你無法計算從未抽到過的顏色——但你**可以**計算只見過一次的顏色，即「單例」。這種稱為**古德–圖靈估計**的技巧認為：單例在全部抽樣中所佔的比例，可以估計仍隱藏在未見顏色中的機率質量。100 次抽取中有 10 次屬於只出現一次的顏色，因此下一顆球是全新顏色的機率約為 **10 / 100 = 10%**。
> 那些只見過一次的顏色不是錯誤。它們是在衡量你自己的無知：許多顏色只出現一次，是樣本在告訴你，世界上還有更多顏色，只是你尚未抽到。
> **現在把顏色換成生日，**把袋子換成模型的訓練文字。假設在模型見過的生日中，**五個中有一個只出現一次**。同樣的技巧表明：大約五分之一的機率質量落在模型實際上從未見過的生日上——而生日沒有可供兜底的規律（你無法推算出某人的生日）。因此，面對一個只見過一次或從未見過的日期，模型就像面對一枚無法判斷偏向的硬幣；大約**五個就會錯一個**。再聰明也無法解決這一點：那裡本來就沒有東西可學。
> 這就是整個論證的縮影：單例率衡量這個世界有多少內容無法從這批資料中學到，並由此形成錯誤的**下限**。這也解釋了模型為何幾乎不會答錯首都——巴黎反覆出現，它的單例率接近於零，因此有充足的資訊可學。

這解釋了幻覺從何而來，卻沒有解釋它們為何會留存下來——為什麼模型經過後來所有旨在讓它更有幫助、更誠實的訓練之後，仍會虛張聲勢，而不是承認不確定。論文在這裡用了一個幾乎令人不適地貼切的類比。設想一個學生在考試中不知道答案。如果留空得零分，而猜測有可能得一分，那麼追求最高分的做法就是猜——要自信、要具體，絕不能說「我不確定」。學生會學會這一點。事實證明，模型也會——因為我們的評分方式完全相同。作者檢查了這個領域真正競爭的基準，也就是模型經過調

校要攀登的排行榜，發現幾乎所有基準都把「我不知道」和錯誤答案同樣計為零分。在這條規則下，總是猜測的模型會擊敗另一個除了會誠實標示不確定性之外完全相同的模型。從相當字面的意義上說，是我們的評分把它們推向了幻覺。



基準測試可能讓猜測成為理性選擇。在多數基準採用的評分方式下（左），錯誤答案和誠實的「我不知道」都得零分——因此猜測只可能有利。若在問題中寫明規則，並對錯誤答案扣分（右），那麼在不確定時放棄作答就可能成為較佳選擇。這改變了測試所獎勵的行為；但它本身並不能解決幻覺。

這部分最值得記住，因為它與常見標題背道而馳。幻覺經常被描述成技術不可避免、近乎神秘的極限。論文在這兩點上都提出異議。預訓練中的錯誤下限並不神秘——它只是普通的統計誤差，是機器學習幾十年來一直理解的那種問題。幻覺的持續存在也並非不可避免——它在一定程度上源於我們設計、也能夠改變的誘因。一個在不確定時乾脆拒絕回答的系統根本不會產生幻覺；已部署模型不這樣做，是因為我們的計分板會懲罰拒答。

作者做了什麼

論文分為三個部分。第一，作者用數學論證說明，預訓練期間某些幻覺在統計上不可避免：他們證明，「只生成有效文字」至少與「這條陳述是否有效」的二元分類問題一樣困難。第二，作者調查十項有影響力的基準，並據此論證，主流的正確率式指標獎勵猜測而不是棄答。第三，作者提出修正方案並用案例研究檢驗它：採用**開放式評分準則**，把評分規則寫在問題本身之中（例如，「答對得 1 分，答錯得 −1 分，因此若把握低於 50% 就應棄答」），讓模型知道何時誠實會得到獎勵。他們在 GoogleGemini 3 Pro、OpenAI GPT-5、xAI Grok 4 和 Anthropic Claude Opus 4.5 四個尖端模型上測

試這一做法，使用SimpleQA的 4,326 個事實問題。作者明確表示，這項案例研究只作說明，「不是跨模型的受控評估」（使用預設設定，沒有調校，也沒有成本正規化）。

他們發現了什麼

- **預訓練必然造成一些錯誤**。模型生成自信虛假陳述的比率存在一個下限，大致是由該模型構建的最佳「這條陳述是否有效」分類器錯誤率的兩倍。對於沒有可學習規律的事實，這個下限至少等於單例率——即訓練中只出現一次的事實所佔比例。即使資料完全乾淨，一些幻覺仍不可避免。
- **評分確實獎勵猜測**。在普通的對錯評分下，從不棄答是最佳策略；作者的調查發現，絕大多數常用基準都把「我不知道」直接算作錯誤。他們自己的一個鮮明例子是：在 SimpleQA 測試上，原始正確率略微偏愛 OpenAI 的 o4-mini——它幾乎每題都答，卻有四分之三以上答錯——而不是 GPT-5-mini；後者會在不確定時棄答，因此犯錯少得多。更冒進的模型反而在計分板上看起來更好。
- **開放式評分準則扭轉了誘因（在這項案例研究中）**。作者測試了一種簡單的幻覺緩解方法：讓模型回答兩次，如果兩個答案不一致就棄答。在標準正確率下，這種方法減少了錯誤，但也降低了正確率——因此指標會阻礙人們採用它。在開放式評分準則下，同一種方法在一系列錯誤懲罰設定中，對四個模型都更佔優勢；原始正確率曾因 GPT-5-mini 在不確定時棄答而懲罰它，但規則公開後，它也領先於 o4-mini（每個模型 n = 4,326 個問題）。

這可能意味著什麼

減少幻覺，主要並不是發明更多針對幻覺的專門測試，而是改變主流基準給不確定性打分的方式，讓承認「我不知道」不再受到懲罰。只要計分板不變，減少幻覺就會繼續讓模型損失正確率分數，從而繼續受到阻礙——這正是作者把問題稱為「社會技術」問題的原因：一部分需要更好的指標，一部分需要讓有影響力的排行榜採納它。

這項研究沒有證明什麼

- 它**沒有**表明開放式評分準則能修復真實環境中的幻覺。支撐這一主張的實驗是一項小型、刻意保持**非受控**的案例研究——四個模型、預設設定、一種選定的緩解方法、一項事實問答測試——目的是展示誘因如何翻轉，而不是給模型排名或證明普遍有效。
- 它**沒有**聲稱評分是唯一原因。訓練資料中的錯誤、真正困難的問題和陌生提示仍是彼此獨立的來源。
- 它**不支持**「幻覺不可避免」這句流行說法。作者的論點恰好相反：一個只回答可核查問題、否則就說「我不知道」的系統永遠不會產生幻覺。
- 它**沒有**讓預訓練的錯誤下限消失——它解釋並界定了這個下限，而且該界限針對的是自信的事實錯誤，不是模型的全部行為。
- 它**沒有**證明開放式評分準則單獨使用就已足夠。它們改變評估所獎勵的行為，卻不能替代檢索、

工具使用或校準更好的模型。

證據有多強？

- 核心證據是**數學**——正式的下界，而不是測量結果。作為理論論證，它在自身假設內成立。
- 它建立在**刻意簡化的問題模型**上；作者自己指出，把每個回答歸為正確、錯誤或「我不知道」是一種「虛假的三分法」，而最清晰的下界也採用了理想化的「任意事實」設定。
- 基準調查是一個**小型、經篩選的樣本**——十項有影響力的評估，不是窮盡式審計。
- 案例研究**真實但有限**：四個尖端模型、一種緩解方法、僅 SimpleQA、預設設定，並明確註明「不是受控評估」。它是誘因論點的概念驗證，不是基準測試結果。
- 還應說明其立場：四位作者中有三位正在或曾經受僱於 **OpenAI**，而論文主張該領域應改變模型評估方式。這是一項由利益相關方提出、論證充分的立場，而不是中立的外部審查——應當權衡，而非因此否定。（值得肯定的是，論文對自家模型 o4-mini 和 GPT-5-mini 的批評，與對其他模型同樣直接。）

為什麼重要

它重新界定了一個被嚴重炒作的問題。「幻覺」通常要麼被包裝成詭異的缺陷，要麼被說成無法跨越的高牆；這篇論文則把它還原為普通現象，而且部分是我們自己造成的——一個可以理解的統計下限，疊加在我們親手選擇的誘因之上。更廣泛的教訓更平實，也更有用：可靠性的進一步提升，可能既取決於我們衡量什麼，也取決於我們建造什麼。

清晰摘要

語言模型自信地給出錯誤答案，來自兩個方面。第一是統計因素：當一個事實沒有可學的規律時，經過語言模仿訓練的模型有時會答錯，而這個下限可以估計（例如依據訓練中有多少事實只出現一次）。第二是誘因因素：幾乎所有用於模型排名的基準，都把「我不知道」和錯誤答案打成同樣的分數，所以猜測總會佔優——甚至一個有四分之三時間答錯的模型，也可能勝過一個會棄答、因而更誠實的模型。作者提出的不是另一種幻覺測試，而是「開放式評分準則」：在問題裡寫明評分方式。在四個尖端模型的案例研究中，這扭轉了誘因，使一種減少幻覺的方法獲得獎勵，而不再受到懲罰。這是一篇結合理論與調查、並附帶一項明確註明非受控小型實驗的論文，已在 *Nature* 接受同行評審；所提方案前景可期，但尚未證明能大規模奏效，而作者認為幻覺既不神秘，也並非嚴格不可避免。

不繞彎核查

論文證明了什麼：預訓練期間某些幻覺不可避免的數學下界（對於沒有規律的事實，至少是「單例率」）；一項調查發現，多數領先基準不給「我不知道」任何分數；以及一項四模型案例研究，其中把評分規則寫進提示（「開放式評分準則」），會讓減少幻覺的方法勝出，而普通正確率原本會懲罰它。

什麼合理但尚未證明：如果主流基準採用開放式評分準則，已部署模型的幻覺會顯著減少。支撐實驗規模很小，而且明確屬於非受控研究。

它沒有證明什麼：幻覺不可避免（它主張相反）；評分是唯一原因；幻覺可以被徹底消除；這項案例研究能給四個模型排出高下。

主要限制：刻意簡化的正確/錯誤/「我不知道」模型（作者稱之為「虛假的三分法」）；小型、經篩選的基準調查（十項評估）；非受控案例研究（四個模型、一種緩解方法、一項測試、預設設定）；以及一項由 OpenAI 主導、討論該領域應如何評估模型的論證。

普通讀者應該有多大信心？可以高度確信幻覺既不神秘，也並非嚴格不可避免，而且主流基準目前確實獎勵猜測。對於所提修正方案能否奏效，應持中等信心——它已有真實的概念驗證，但尚未證明可以廣泛、大規模地發揮作用。

來源

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