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 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

# 為什麼模型會猜，以及我們為什麼教會它們這麼做

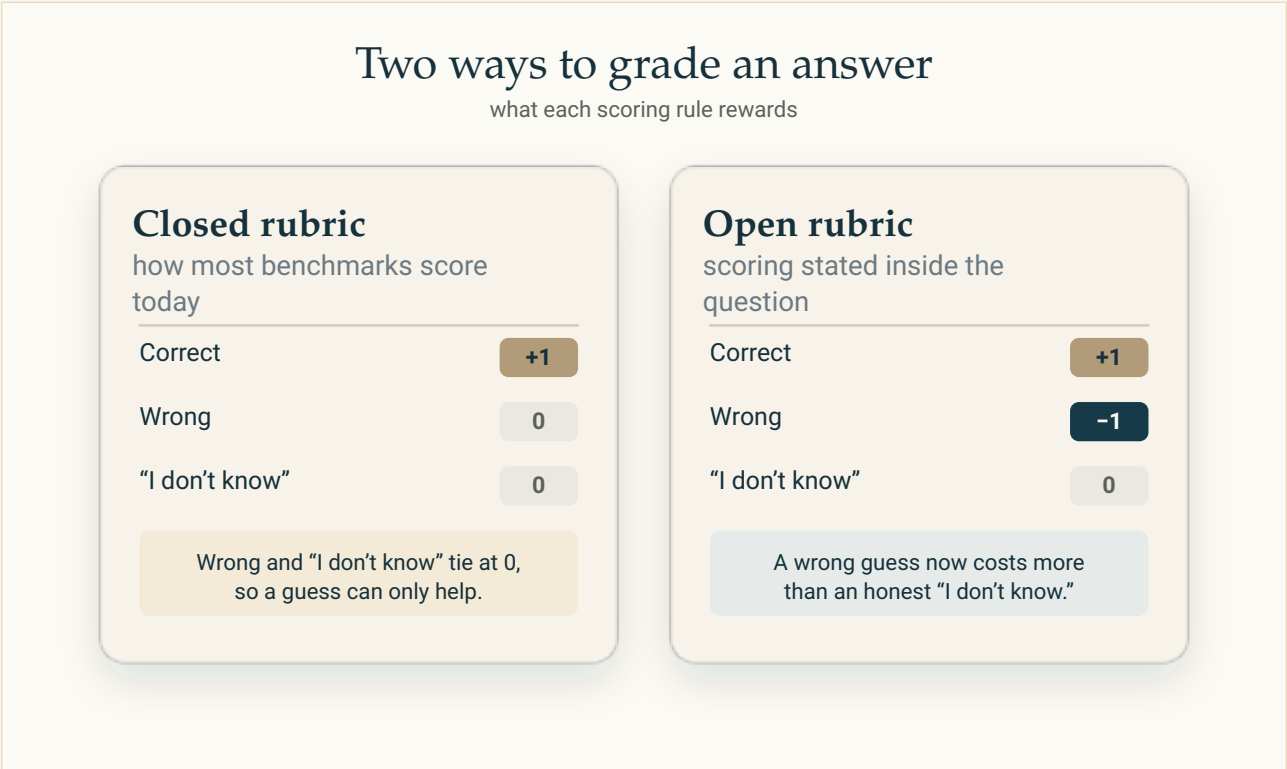
要求一個大型語言模型說出陌生人的生日，它可能會以像是照著卡片念一樣篤定的口吻回答「3 月 7 日」——但連續問三次，三次都答錯，而且每次的日期都不同。作者正是以這類例子開場：向領先模型提出簡單的事實問題——某人的生日，或某個冷門縮寫代表什麼——每個模型都自信地編造出不同答案，而且沒有一個正確。業界稱這種現象為幻覺，聽起來彷彿是知覺出了問題。這篇論文的第一步，就是揭開它的神祕面紗。

先從模型的建造方式開始。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，實際上學的是流暢語言的樣子。現在考慮一個背後沒有模式可循的事實——某個特定人的生日。如果那個日期在訓練文字中只出現過一次，甚至從未出現，對這個模式學習者而言就沒有任何可供掌握的線索：從模型的角度看，答案是任意的。作者借用一個更早的想法（來自 Alan Turing，但原本用於不同問題）把這件事說得精確：如果五分之一的生日在資料中只出現一次，就應該預期模型至少會答錯其中五分之一——不是因為模型壞了，而是因為根本沒有任何東西可學。（同樣的道理，模型幾乎不會答錯某國的首都：這些資訊會不斷出現。）作者謹慎地指出，分辨一個真實陳述和一個貌似合理的錯誤陳述本身就是難題，而只產生真實陳述至少同樣困難。錯誤的下限早已內建其中。

> **背後的想法：如何計算你尚未見過的事物**
>
> 這建立在一個真正巧妙的想法上——它比語言模型更早出現，而且值得好好認識。
> **先從一袋彩色球開始**。你不知道袋子裡有多少種顏色。你逐一抽取 100 顆並記錄：紅色 40 顆、藍色 25 顆、綠色 15 顆、黃色 5 顆、紫色 3 顆、橘色 2 顆——接著是**十種各自恰好出現一次的不同顏色**。
> 現在來看 Turing 實際面對的問題——雖然是完全不同的問題：下一顆球是你完全沒見過的顏色的機率是多少？你無法計算自己從未抽到過的事物——但你**可以**計算那些恰好只見過一次的顏色，也就是「單例」（singletons）。這個稱為 **Good–Turing 估計**的技巧是：抽取結果中單例所占的比例，可以估計尚未見過的顏色所占的機率。100 次抽取中有 10 次是只出現一次的顏色，所以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。它們量測的是你自己的無知：許多顏色只出現一次，正是樣本在告訴你，這個世界還有更多你尚未抽到的顏色。
> **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，**五分之一恰好只出現一次**。同樣的技巧告訴我們：大約五分之一的機率落在模型實際上從未見過的生日上——而生日沒有可依循的模式（你無法推算出某人的生日）。因此，對於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正反面機率的硬幣，大約會答錯其中**五分之一**。再聰明也無法修正這件事：根本沒有任何東西可學。
> 這就是整個論證的縮影：單例率量測了從這份資料中無法學會的世界比例，而這會成為錯誤的**下限**。這也解釋了為什麼模型幾乎不會答錯首都——*Paris* 不斷出現，它的單例率接近零，因此有大量資訊可學。

這說明了幻覺從何而來，但沒有解釋它們為什麼持續存在——為什麼模型經過所有旨在讓它們更有用、更誠實的後續訓練後，仍然裝懂，而不是承認自己不確定。這裡，論文的比喻幾乎貼切得令人不自在。想像一名學生在考試中不知道答案。如果空白作答得 0 分，而猜測可能得 1 分，那麼以最大化分數為目標的做法就是猜——自信、具體，絕不說「我不確定」。學生會學到這一點。結果模型也會，因為我們用同樣的方式評分它們。作者檢視了這個領域實際競爭的基準測試，也就是模型被調整來攀

登的排行榜，發現其中幾乎全部都把「我不知道」和錯誤答案給予完全相同的分數：0 分。在這條規則下，一個永遠猜測的模型會勝過另一個原本相同、但誠實標示不確定性的模型。嚴格說來，是我們用評分把模型塑造成這樣。




基準測試可能讓猜測變得合理。在大多數基準測試採用的評分方式（左）下，錯誤答案和誠實的「我不知道」都得 0 分——因此猜測只可能帶來好處。若在題目中說明規則，並對答錯施加懲罰（右），不確定時選擇不作答就可能成為更好的做法。這會改變測試獎勵的行為；但它本身並不能解決幻覺。



這是值得記住的重點，因為它與常見的標題論調相反。幻覺常被描述成技術上不可避免、近乎神祕的限制。這篇論文在兩方面都提出異議。預訓練的下限並不神祕——它就是普通的統計誤差，是機器學習幾十年來早已理解的那一類誤差。而幻覺的持續存在也不是不可避免的——它部分源自我們建立的誘因，而我們可以改變它。如果系統在不確定時直接拒絕回答，它根本不會產生幻覺；已部署的模型之所以不這麼做，是因為我們的排行榜會懲罰拒答。

# 作者做了什麼

這篇論文分為三個部分。第一部分是數學論證：透過展示「只生成有效文字」至少和二元的「這個陳述是否有效？」分類問題一樣困難，說明預訓練期間某些幻覺在統計上必然發生。第二部分是論證，並以對十個具影響力的基準測試所做的調查作為支持：主流的準確率式指標獎勵猜測，而不是不作答。第三部分提出一種修正方案並進行個案研究來測試：**開放式評分規則**（open-rubric）評估，也就是把評分方式直接寫在題目中（例如「答對得 1 分，答錯得 −1 分，所以若你的確定程度低於 50% 就不作答」），讓模型能辨識何時誠實會得到獎勵。他們用四個前沿模型測試這個方法——Google 的 Gemini

3 Pro、OpenAI 的 GPT-5、xAI 的 Grok 4，以及 Anthropic 的 Claude Opus 4.5——使用 SimpleQA 的 4,326 道事實問題。他們明確表示，這項個案研究只是示範，「不是跨模型的受控評估」（使用預設設定、未調整、未進行成本標準化）。

# 他們發現了什麼

- **預訓練必然造成某些錯誤**。模型產生自信錯誤陳述的比率，其下限約為以該模型建立的最佳「這個陳述是否有效？」分類器錯誤率的兩倍。對於沒有可學習模式的事實，這個下限至少是單例率——訓練資料中恰好只出現一次的事實所占比例。即使資料完全乾淨，某些幻覺仍然不可避免。
- **評分確實獎勵猜測**。在一般的答對／答錯評分下，永不拒答是最佳策略；作者的調查發現，絕大多數熱門基準測試都把「我不知道」直接算作錯誤。他們自身的研究提供了一個鮮明例子：在 SimpleQA 測試中，原始準確率稍微偏向 OpenAI 的 o4-mini——它幾乎回答所有問題，而且超過四分之三的回答是錯的——而不是 GPT-5-mini；後者在不確定時會拒答，因此犯錯少得多。更魯莽的模型在排行榜上看起來更好。
- **開放式評分規則會翻轉誘因（在他們的個案研究中）**。他們測試一種簡單的幻覺緩解方法（讓模型回答兩次，若兩個答案不一致就拒答）。在標準準確率下，這種緩解方法會減少錯誤，但也會降低準確率——因此該指標不鼓勵採用它。在開放式評分規則下，同一種緩解方法在一系列懲罰設定中對四個模型都更有利；而 GPT-5-mini——原始準確率曾因它在不確定時拒答而受到懲罰——在公開說明評分方式後，表現勝過 o4-mini（每個模型 n = 4,326 道問題）。

# 這可能意味著什麼

降低幻覺，大多不是再發明更多專門測試幻覺的方法，而是改變主流基準測試評分不確定性的方式，讓承認「我不知道」不再受到懲罰。在排行榜改變之前，降低幻覺會持續讓模型付出準確率分數的代價，也就持續受到抑制——這正是作者把問題框定為「社會—技術性」（socio-technical）的原因：一部分是更好的指標，一部分是促使有影響力的排行榜採用它。

# 這篇論文沒有證明什麼

- 它**沒有**顯示開放式評分規則能在真實環境中解決幻覺。支持性實驗是一項規模小、刻意**不受控**的個案研究——四個採用預設設定的模型、一種選定的緩解方法、一項事實問答測試——目的是展示誘因翻轉，不是替模型排名，也不是證明普遍有效。
- 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，以及不熟悉的提示，仍然是獨立的來源。
- 它**不**支持「幻覺不可避免」這句流行說法。作者主張相反的觀點：一個只回答可查核問題、其他情況都說「我不知道」的系統，永遠不會產生幻覺。
- 它**不會**讓預訓練下限消失——它解釋並界定了這個下限，而該界限針對的是自信的事實錯誤，不是模型的所有行為。

- 它**沒有**顯示開放式評分規則單獨就足夠。它們改變的是評估所獎勵的行為；它們不能取代檢索、工具使用或校準更好的模型。

# 證據有多強？

- 核心是**數學**——是形式化的下界，而不是測量結果。作為理論論證，它在自身的前提下是健全的。
- 它建立在對問題**刻意簡化的模型**上；作者自己也指出，把每個回答都視為正確、錯誤或「我不知道」的「錯誤三分法」，以及為了得到最簡潔下界而使用的理想化「任意事實」設定。
- 基準測試調查是**小型且經過挑選的樣本**——十個具影響力的評估，而不是詳盡的稽核。
- 個案研究**真實但有限**：四個前沿模型、一種緩解方法、只有 SimpleQA、預設設定，並明確表示「不是受控評估」。它是對誘因論證的概念驗證，不是基準測試結果。
- 值得說明觀察角度：四位作者中有三位目前或曾任職於 **OpenAI**，而論文主張這個領域應該改變評估模型的方式。這是出自利益相關方、論證充分的立場，不是中立的外部審查——應該納入權衡，而不是因此否定。（值得肯定的是，論文同樣直接把批評指向自家的 o4-mini 和 GPT-5-mini，也同樣指向其他模型。）

# 為什麼重要

它重新框定了一個被大肆炒作的問題。「幻覺」往往被描述成詭異的缺陷，或一道不可撼動的高牆；這篇論文則讓它變得平凡，而且部分是我們自找的——一個可以真正理解的統計下限，疊加在我們自己選擇的誘因之上。更廣泛、也更實用的啟示是：可靠性的進一步提升，可能既取決於我們測量什麼，也取決於我們建造什麼。

# 重點摘要

語言模型自信的錯誤回答來自兩個地方。第一個是統計因素：當一個事實沒有可學習的模式時，受訓來模仿語言的模型有時會答錯，而這個下限可以估計（例如，根據訓練資料中只出現一次的事實數量）。第二個是誘因：模型用來排名的幾乎每個基準測試，都把「我不知道」算得和錯誤答案一樣，所以猜測永遠勝出——甚至一個四分之三的回答都錯的模型，也能勝過一個會拒答、因而更誠實的模型。作者提出的不是另一個幻覺測試，而是「開放式評分規則」：把評分方式寫在題目中。在四個前沿模型上的個案研究中，這會翻轉誘因，使降低幻覺的方法受到獎勵，而不是懲罰。這是一篇結合理論與調查的論文，附有一項規模小且明確不受控的實驗，並已在 *Nature* 同行評審；這個修正方案很有希望，但尚未證明能大規模運作，而作者主張幻覺既不神祕，也不是嚴格意義上不可避免的。

# 務實檢視

**論文顯示了什麼：**一個數學下界，說明預訓練期間某些幻覺必然發生（對沒有模式的事實而言，至少是「單例率」）；一項調查發現，多數領先基準測試不會為「我不知道」給分；以及一項四模型個案研究，在其中，於提示中說明評分方式（「開放式評分規則」）會讓降低幻覺的方法勝出，而普通準確率原本會懲罰它。

**合理但尚未證明的事：**把開放式評分規則納入主流基準測試，可能會在已部署的模型中有意義地降低幻覺。支持性實驗規模小，而且明確不受控。

**它沒有顯示什麼：**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；個案研究為四個模型彼此排名。

**主要限制：**刻意簡化的正確／錯誤／「我不知道」模型（作者稱之為「錯誤三分法」）；小型且經過挑選的基準測試調查（十項評估）；不受控的個案研究（四個模型、一種緩解方法、一項測試、預設設定）；以及一篇由 OpenAI 主導、主張這個領域應如何評估模型的論證。

**一般讀者應有多大信心？**對於幻覺既不神祕也不是嚴格意義上不可避免，以及主流基準測試目前確實獎勵猜測，可以有高度信心。對於提出的修正方案有所幫助，則應抱持中等信心——它現在有了真實的概念驗證，但尚未證明能廣泛且大規模地運作。

---

# 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