The Clean Paper

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

Vì sao mô hình ngôn ngữ "ảo giác" — và vì sao cách chúng ta chấm điểm khiến nó tiếp tục

Những phát biểu sai nhưng đầy tự tin mà ta gọi là "ảo giác" không phải một trục trặc bí ẩn: một phần là hệ quả thống kê của huấn luyện, và chúng tồn tại dai dẳng vì benchmark phổ biến thưởng cho một câu đoán tự tin hơn một lời trung thực "tôi không biết". Một nghiên cứu tình huống trên bốn mô hình frontier cho thấy việc nêu quy tắc chấm ngay trong prompt ("open rubric") có thể đảo chiều động lực này.

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

Vì sao mô hình đoán mò, và vì sao chính chúng ta dạy chúng làm vậy

Hỏi một mô hình ngôn ngữ lớn ngày sinh của một người xa lạ, nó có thể trả lời "ngày 7 tháng 3" với vẻ chắc chắn như đang đọc từ một tấm thẻ — rồi sai, ba lần liên tiếp, với ba ngày khác nhau. Các tác giả đưa ra đúng kiểu ví dụ này: những mô hình hàng đầu được hỏi một câu hỏi thực tế đơn giản — ngày sinh của một người, hay một chữ viết tắt hiếm nghĩa là gì — và mỗi mô hình đều tự tin bịa ra một đáp án khác nhau, không cái nào đúng. Ngành công nghiệp gọi hiện tượng này là *hallucination* (ảo giác), một cái tên khiến nó nghe như lỗi nhận thức. Bước đầu tiên của bài báo là gỡ bỏ vẻ bí ẩn đó.

Hãy bắt đầu từ cách một mô hình được tạo ra. Trong giai đoạn huấn luyện đầu tiên và lớn nhất, về cơ bản nó học xem ngôn ngữ trôi chảy trông như thế nào bằng cách đọc một lượng văn bản khổng lồ. Giờ hãy lấy một sự thật không có quy luật phía sau — ngày sinh của một người cụ thể. Nếu ngày đó chỉ xuất hiện một lần trong dữ liệu huấn luyện, hoặc không hề xuất hiện, thì một hệ thống học mẫu không có gì để bám vào: từ góc nhìn của mô hình, đáp án là tùy ý. Các tác giả làm ý này trở nên chính xác bằng cách mượn một ý tưởng cũ của Alan Turing, vốn xuất phát từ một bài toán khác: nếu cứ năm ngày sinh thì có một ngày chỉ xuất hiện đúng một lần trong dữ liệu, ta nên kỳ vọng mô hình sai ít nhất một phần năm trong số đó — không phải vì mô hình "hỏng", mà vì đơn giản là chưa từng có đủ thông tin để học. (Theo cùng logic, mô hình hầu như không sai thủ đô của một quốc gia: những dữ kiện ấy xuất hiện liên tục.) Các tác giả lập luận khá cẩn thận rằng việc phân biệt một phát biểu đúng với một phát biểu sai nhưng nghe hợp lý tự nó đã là bài toán khó, và việc tạo ra *chỉ* các phát biểu đúng ít nhất cũng khó bằng vậy. Một mức sai số sàn đã được cài vào bài toán ngay từ đầu.

> **Ý tưởng bên dưới: làm sao đếm thứ bạn chưa từng thấy**
>
> Điều này dựa trên một ý tưởng thật sự thông minh — có trước các mô hình ngôn ngữ rất lâu, và đáng để hiểu đúng.
> **Hãy bắt đầu với một túi bi nhiều màu.** Bạn không biết có bao nhiêu màu trong túi. Bạn rút 100 viên, từng viên một, rồi đếm: đỏ 40, xanh lam 25, xanh lá 15, vàng 5, tím 3, cam 2 — rồi **mười màu khác nhau, mỗi màu chỉ xuất hiện đúng một lần.**
> Câu hỏi Turing thực sự gặp, trong một bài toán hoàn toàn khác, là: *xác suất viên tiếp theo có một màu bạn chưa từng thấy là bao nhiêu?* Bạn không thể đếm thứ chưa từng rút ra — nhưng bạn **có thể** đếm những màu chỉ thấy đúng một lần, các "singleton". Mẹo có tên **ước lượng Good–Turing** nói rằng phần mẫu thuộc các singleton ước lượng phần xác suất vẫn đang ẩn trong những màu bạn chưa nhìn thấy. Mười trong một trăm lần rút là các màu xuất hiện một lần, vì vậy xác suất viên tiếp theo có màu hoàn toàn mới vào khoảng **10 / 100 = 10%.**
> Những màu chỉ thấy một lần đó không phải là *sai sót*. Chúng là phép đo sự thiếu hiểu biết của chính bạn: nhiều màu xuất hiện đúng một lần là cách mẫu dữ liệu nói rằng thế giới còn nhiều màu khác mà bạn đơn giản chưa rút trúng.
> **Giờ đổi màu thành ngày sinh,** và đổi túi bi thành văn bản huấn luyện của mô hình.

> Giả sử trong số các ngày sinh mô hình từng thấy, **một phần năm chỉ xuất hiện đúng một lần.** Dùng cùng mẹo: khoảng một phần năm khối xác suất nằm ở những ngày sinh mà trên thực tế mô hình gần như chưa từng thấy — và ngày sinh không có quy luật nào để suy ra (bạn không thể *tính* ngày sinh của một người). Vì vậy một ngày chỉ thấy một lần, hoặc chưa từng thấy, là một đồng xu mà mô hình không biết phải nghiêng về phía nào; nó sẽ sai khoảng **một trên năm** trường hợp. Không có sự "thông minh" nào sửa được chuyện này: dữ liệu cần để học chưa từng tồn tại.
> Đó là toàn bộ lập luận thu nhỏ: tỷ lệ singleton đo xem bao nhiêu phần của thế giới là không thể học được *từ bộ dữ liệu này*, và nó trở thành một **mức sàn** dưới tỷ lệ lỗi. Đây cũng là lý do mô hình gần như không bao giờ bỏ lỡ thủ đô — *Paris* xuất hiện liên tục, tỷ lệ singleton gần bằng không, nên có rất nhiều tín hiệu để học.

Điều đó giải thích ảo giác bắt nguồn từ đâu. Nhưng nó chưa giải thích vì sao chúng *tồn tại dai dẳng* — vì sao sau tất cả những giai đoạn huấn luyện sau này nhằm làm mô hình hữu ích và trung thực hơn, nó vẫn nói liều thay vì thừa nhận không chắc chắn. Ở đây phép so sánh của bài báo gần như khó chịu vì quá đúng. Hãy hình dung một học sinh trong kỳ thi không biết đáp án. Nếu bỏ trống được 0 điểm và đoán có thể được 1 điểm, chiến lược tối đa hóa điểm là đoán — thật chắc chắn, thật cụ thể, tuyệt đối không nói "em không chắc". Học sinh học được điều này. Mô hình cũng vậy — vì chúng ta chấm chúng theo đúng cách đó. Các tác giả xem xét những benchmark mà lĩnh vực thực sự cạnh tranh, những bảng xếp hạng mà mô hình được tinh chỉnh để leo lên, và thấy gần như tất cả đều cho "tôi không biết" cùng điểm với một câu trả lời sai: 0. Dưới quy tắc ấy, một mô hình luôn đoán sẽ đánh bại một mô hình giống hệt nhưng trung thực báo rằng nó không chắc. Theo nghĩa khá sát chữ, chính cách chấm điểm của chúng ta huấn luyện chúng làm thế.

Two ways to grade an answer

what each scoring rule rewards

Closed rubric

how most benchmarks score today

Correct	+1
Wrong	0
"I don't know"	0

Wrong and "I don't know" tie at 0, so a guess can only help.

Open rubric

scoring stated inside the question

Correct	+1
Wrong	−1
"I don't know"	0

A wrong guess now costs more than an honest "I don't know."

Benchmark có thể khiến đoán mò trở thành lựa chọn hợp lý. Với cách chấm phổ biến ở hầu hết benchmark (trái), một câu sai và một câu trung thực "tôi không biết" đều được 0 điểm — vì thế đoán chỉ có thể có lợi. Nếu quy tắc được ghi ngay trong câu hỏi và trả lời sai bị phạt (phải), thì khi không chắc chắn, từ chối trả lời có thể trở thành lựa chọn tốt hơn. Điều này thay đổi điều mà bài kiểm tra thưởng; tự nó không giải quyết ảo giác.

Original diagram — The Clean Paper CC BY 4.0

Đây là phần đáng giữ lại, vì nó đi ngược với tiêu đề thường thấy. Ảo giác hay được mô tả như một giới hạn tất yếu, gần như huyền bí của công nghệ. Bài báo phản đối cả hai điều đó. Mức sàn từ tiền huấn luyện không bí ẩn — đó là sai số thống kê bình thường, loại mà học máy đã hiểu hàng chục năm. Và việc lỗi tiếp tục tồn tại không phải hoàn toàn tất yếu — một phần là do động lực mà chúng ta tự xây và có thể thay đổi. Một hệ thống đơn giản từ chối trả lời khi không chắc sẽ không tạo ra phát biểu sai tự tin; lý do các mô hình triển khai không cư xử như vậy là vì bảng điểm của chúng ta phạt sự từ chối.

Các tác giả đã làm gì

Bài báo có ba phần. Thứ nhất là một lập luận toán học rằng một phần ảo giác bị ép buộc về mặt thống kê trong giai đoạn tiền huấn luyện, bằng cách cho thấy bài toán "chỉ sinh văn bản hợp lệ" ít nhất cũng khó bằng bài toán phân loại nhị phân "phát biểu này có hợp lệ không?". Thứ hai là lập luận — được hỗ trợ bởi khảo sát mười benchmark có ảnh hưởng — rằng các chỉ số kiểu accuracy phổ biến thưởng cho việc đoán hơn là từ chối. Thứ ba là một cách sửa được đề xuất và một nghiên cứu tình huống kiểm tra nó: đánh giá với **open rubric**, trong đó cách chấm điểm được nêu ngay trong câu hỏi (ví dụ "đúng được 1 điểm, sai bị −1, vì vậy hãy từ chối nếu bạn chắc dưới 50%"), để mô hình biết khi nào sự trung thực được thưởng. Họ thử điều này trên bốn mô hình frontier — Gemini 3 Pro của Google, GPT-5 của OpenAI, Grok 4 của xAI và Claude Opus 4.5 của Anthropic — dùng 4.326 câu hỏi thực tế của SimpleQA. Họ nói rất rõ đây chỉ là minh họa, "không phải một đánh giá có kiểm soát giữa các mô hình" (cài đặt mặc định, không tinh chỉnh, không chuẩn hóa chi phí).

Họ tìm thấy gì

- **Tiền huấn luyện buộc phải có một phần lỗi.** Tỷ lệ mô hình tạo ra phát biểu sai một cách tự tin có một cận dưới liên hệ với (xấp xỉ gấp đôi) tỷ lệ lỗi của bộ phân loại tốt nhất có thể xây từ nó cho câu hỏi "phát biểu này có hợp lệ không?". Với các dữ kiện không có mẫu để học, mức sàn đó ít nhất bằng *tỷ lệ singleton* — phần các dữ kiện chỉ xuất hiện đúng một lần trong dữ liệu huấn luyện. Một phần ảo giác là không thể tránh ngay cả với dữ liệu hoàn toàn sạch.
- **Cách chấm cụ thể thưởng cho việc đoán.** Với chấm đúng/sai thông thường, không bao giờ từ chối là chiến lược tối ưu, và khảo sát của các tác giả cho thấy phần lớn benchmark phổ biến chấm "tôi không biết" đơn giản là sai. Một ví dụ nổi bật ngay từ các mô hình OpenAI: trên SimpleQA, accuracy thô hơi *ưu ái* o4-mini — mô hình trả lời gần như mọi câu

và sai hơn ba phần tư số lần — hơn GPT-5-mini, vốn mắc ít lỗi hơn nhiều vì biết từ chối khi không chắc. Mô hình liều lĩnh hơn trông tốt hơn trên bảng điểm.
- **Open rubric đảo chiều động lực (trong nghiên cứu tình huống).** Họ thử một biện pháp giảm ảo giác đơn giản (cho mô hình trả lời hai lần và từ chối nếu hai câu không khớp). Theo accuracy chuẩn, biện pháp này giảm lỗi *nhưng cũng giảm accuracy* — vì vậy chỉ số lại khuyến khích không dùng nó. Với open rubric, cùng biện pháp đó thắng ở cả bốn mô hình trong một dải mức phạt; và GPT-5-mini — vốn bị accuracy thô phạt vì từ chối khi không chắc — vượt o4-mini khi quy tắc chấm được nêu công khai (n = 4.326 câu hỏi cho mỗi mô hình).

Điều này có thể có ý nghĩa gì

Giảm ảo giác chủ yếu không phải là phát minh thêm những bài kiểm tra chuyên biệt về ảo giác. Vấn đề là thay đổi cách các benchmark chính chấm sự không chắc chắn, để thừa nhận "tôi không biết" không còn bị phạt. Chừng nào bảng điểm chưa đổi, giảm ảo giác vẫn làm mô hình mất điểm accuracy và vì thế tiếp tục bị nản lòng — đó là lý do các tác giả gọi đây là vấn đề "xã hội-kỹ thuật": một phần là chỉ số tốt hơn, một phần là thuyết phục các bảng xếp hạng có ảnh hưởng sử dụng nó.

Điều này *không* chứng minh gì

- Nó **không** cho thấy open rubric giải quyết ảo giác ngoài đời thực. Thí nghiệm hỗ trợ là một nghiên cứu tình huống nhỏ và cố ý **không kiểm soát** — bốn mô hình ở cài đặt mặc định, một biện pháp giảm ảo giác, một bài kiểm tra hỏi đáp thực tế — nhằm minh họa *sự đảo chiều động lực*, không phải xếp hạng mô hình hay chứng minh hiệu quả tổng quát.
- Nó **không** khẳng định chấm điểm là *nguyên nhân duy nhất*. Dữ liệu huấn luyện sai, các bài toán thực sự khó và prompt lạ vẫn là các nguồn lỗi riêng.
- Nó **không** ủng hộ câu nói phổ biến rằng ảo giác là tất yếu. Các tác giả lập luận ngược lại: một hệ thống chỉ trả lời các câu có thể kiểm chứng và nói "tôi không biết" trong các trường hợp còn lại sẽ không tạo phát biểu sai tự tin.
- Nó **không** làm mức sàn ở tiền huấn luyện biến mất — nó giải thích và đặt cận cho mức sàn đó, và cận này liên quan đến lỗi dữ kiện tự tin chứ không phải mọi hành vi của mô hình.
- Nó **không** cho thấy open rubric tự nó là *đủ*. Nó thay đổi điều đánh giá thưởng; nó không thay thế retrieval, sử dụng công cụ hay mô hình được hiệu chỉnh xác suất tốt hơn.

Bằng chứng mạnh đến đâu?

- Phần lõi là **toán học** — các cận dưới hình thức, không phải phép đo thực nghiệm. Với tư cách một lập luận lý thuyết, nó đứng vững theo chính các giả định của nó.
- Nó dựa trên những **mô hình hóa cố ý đơn giản hóa**; chính tác giả lưu ý "tam phân giả" khi coi mọi phản hồi là đúng, sai hoặc "tôi không biết", cùng bối cảnh lý tưởng hóa của các "dữ kiện tùy ý" dùng để có cận rõ nhất.
- Khảo sát benchmark là một **mẫu nhỏ được chọn lọc** — mười đánh giá có ảnh hưởng, không phải kiểm toán toàn diện.
- Nghiên cứu tình huống là **dữ liệu thật nhưng hạn chế**: bốn mô hình frontier, một biện pháp giảm ảo giác, chỉ SimpleQA, cài đặt mặc định, và được mô tả rõ là "không phải đánh giá có kiểm soát". Nó là bằng chứng khái niệm cho lập luận về động lực, không phải kết quả benchmark để xếp hạng.
- Cũng nên nói rõ góc nhìn: ba trong bốn tác giả đang hoặc từng làm việc tại **OpenAI**, và bài báo đề xuất toàn lĩnh vực thay đổi cách đánh giá mô hình. Đây là một lập luận có lý lẽ từ một bên có lợi ích liên quan, không phải đánh giá trung lập bên ngoài — cần cân nhắc, không cần gạt bỏ. (Điểm đáng ghi nhận là bài báo phê bình cả các mô hình của chính OpenAI, o4-mini và GPT-5-mini, cũng thẳng tay như với các mô hình khác.)

Vì sao điều này quan trọng

Nó đặt lại khung cho một vấn đề bị thổi phồng rất nhiều. "Ảo giác" thường được bán như một lỗi ma quái hoặc một bức tường không thể vượt qua; bài báo biến nó thành thứ bình thường và một phần do chính chúng ta gây ra — một mức sàn thống kê có thể hiểu được, chồng lên một hệ động lực do con người lựa chọn. Bài học rộng hơn yên tĩnh nhưng hữu ích hơn: tiến bộ tiếp theo về độ tin cậy có thể phụ thuộc vào *thứ chúng ta đo* nhiều không kém thứ chúng ta xây.

Tóm tắt gọn

Các câu trả lời sai nhưng tự tin của mô hình ngôn ngữ đến từ hai nguồn. Thứ nhất là thống kê: khi một dữ kiện không có mẫu để học, mô hình được huấn luyện để bắt chước ngôn ngữ đôi khi sẽ sai, và mức sàn đó có thể được ước tính (chẳng hạn từ số dữ kiện chỉ xuất hiện một lần trong huấn luyện). Thứ hai là động lực: gần như mọi benchmark dùng để xếp hạng mô hình đều chấm "tôi không biết" giống hệt một câu sai, vì vậy đoán luôn có lợi — đến mức một mô hình sai ba phần tư số lần có thể được điểm cao hơn một mô hình trung thực hơn biết từ chối. Đề xuất của các tác giả không phải thêm một bài kiểm tra ảo giác mà là "open rubric": ghi cách chấm ngay trong câu hỏi. Trong một nghiên cứu tình huống trên bốn mô hình frontier, điều này đảo động lực để một phương pháp giảm ảo giác được thưởng thay vì bị phạt. Đây là bài báo kết hợp lý thuyết và khảo sát với một thí nghiệm nhỏ, được nói rõ là không kiểm soát, đã

qua bình duyệt ở *Nature*; cách sửa này hứa hẹn nhưng chưa được chứng minh ở quy mô lớn, và ảo giác được lập luận là không bí ẩn cũng không tuyệt đối tất yếu.

Đánh giá thẳng thắn

Bài báo cho thấy gì: Một cận dưới toán học theo đó một phần ảo giác bị ép buộc trong tiền huấn luyện (ít nhất bằng "tỷ lệ singleton" với các dữ kiện không có mẫu); một khảo sát cho thấy đa số benchmark hàng đầu không cho "tôi không biết" điểm nào; và một nghiên cứu tình huống bốn mô hình trong đó việc ghi rõ cách chấm trong prompt ("open rubric") khiến một phương pháp giảm ảo giác thắng, trong khi accuracy thông thường lại phạt nó.

Điều hợp lý nhưng chưa được chứng minh: Việc đưa open rubric vào các benchmark chính sẽ làm giảm đáng kể ảo giác trong mô hình triển khai. Thí nghiệm hỗ trợ nhỏ và được nói rõ là không kiểm soát.

Điều nó không cho thấy: Ảo giác là tất yếu (bài báo lập luận ngược lại); chấm điểm là nguyên nhân duy nhất; ảo giác có thể bị xóa sạch hoàn toàn; hoặc nghiên cứu tình huống xếp hạng bốn mô hình với nhau.

Hạn chế chính: Mô hình đúng/sai/"tôi không biết" cố ý đơn giản hóa (tác giả gọi là một "tam phân giả"); khảo sát benchmark nhỏ và được chọn lọc (mười đánh giá); nghiên cứu tình huống không kiểm soát (bốn mô hình, một biện pháp, một bài test, cài đặt mặc định); và đây là một lập luận do các tác giả có liên hệ mạnh với OpenAI đưa ra về cách toàn lĩnh vực nên đánh giá mô hình.

Một độc giả phổ thông nên tự tin đến mức nào? Có thể tự tin cao rằng ảo giác không phải hiện tượng bí ẩn hay tuyệt đối tất yếu, và benchmark chính hiện nay thực sự thưởng cho việc đoán. Mức tin vừa phải rằng cách sửa đề xuất sẽ giúp: đã có bằng chứng khái niệm thực tế, nhưng chưa có chứng minh rằng nó hiệu quả rộng và ở quy mô lớn.

Nguồn

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