The Clean Paper

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

言語モデルはなぜハルシネーションするのか —— そして私たちの採点方法がなぜそれを温存するのか

「ハルシネーション」と呼ばれる自信満々の誤情報は、不可解な故障ではない。一部は学習に伴う統計的な副産物であり、さらに主流ベンチマークが正直な「分かりません」より自信ある推測を報いるため残り続ける。四つのフロンティアモデルを使ったケーススタディーでは、採点規則をプロンプト内に明記する「オープン・ルーブリック」がそのインセンティブを反転させた。

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

なぜモデルは当てずっぽうをするのか、そして私たちはなぜそうさせてきたのか

大規模言語モデルに、見知らぬ人の誕生日を尋ねてみる。すると「3 月 7 日です」と、まるでカードを読み上げているかのような落ち着いた自信で答えるかもしれない —— しかも間違っていて、三回聞けば三つの違う日付を返すことすらある。著者たちはまさにこの種の例を示している。著名なモデルに、ある人物の誕生日や、ほとんど知られていない略語の意味といった単純な事実問題を尋ねると、それぞれが自信満々に異なる答えを作り出し、しかもどれも正しくない。業界ではこれをハルシネーション（幻覚）と呼ぶが、その言葉は知覚の故障のように響く。この論文がまず行うのは、その神秘性を取り除くことだ。

まず、モデルがどう作られるかから考える。最初で最大の学習段階では、膨大な量の文章を読むことで、実質的に「流暢な言語とはどのようなものか」を学ぶ。ここで、背後に規則性のない事実 —— たとえば特定の一人の誕生日 —— を考える。その日付が学習テキストに一度しか出ていない、あるいは一度も出ていないなら、パターンを学ぶ仕組みがつかめるものはない。モデルから見れば答えは任意なのだ。著者たちは、別の問題のためにアラン・チューリングが使った古い考え方を借りて、これを厳密化する。もし誕生日の 5 件に 1 件がデータ中に一度しか現れないなら、モデルは少なくとも 5 件に 1 件を間違えると期待すべきだ、というのだ。壊れているからではなく、そもそも学べるものがそこになかったからである。（同じ理屈で、モデルは国の首都をほとんど間違えない。そうした情報は何度も何度も現れるからだ。）著者たちは慎重に、真の文ともっともらしい偽の文を見分けること自体が難しい問題であり、真の文だけを生成することは少なくともそれと同程度に難しい、と論じる。一定の誤りの下限は、最初から組み込まれている。

背後にある考え方：まだ見ていないものをどう数えるか

これは本当に巧妙な考え方に基づいている。言語モデルよりずっと古く、一度きちんと知っておく価値がある。

色のついた球が入った袋を考えよう。何色あるのかは分からない。球を 100 個、一つずつ取り出して数えると、赤 40、青 25、緑 15、黄 5、紫 3、オレンジ 2—— そしてさらに、**それぞれ一度だけ出た異なる色が 10 色**あったとする。

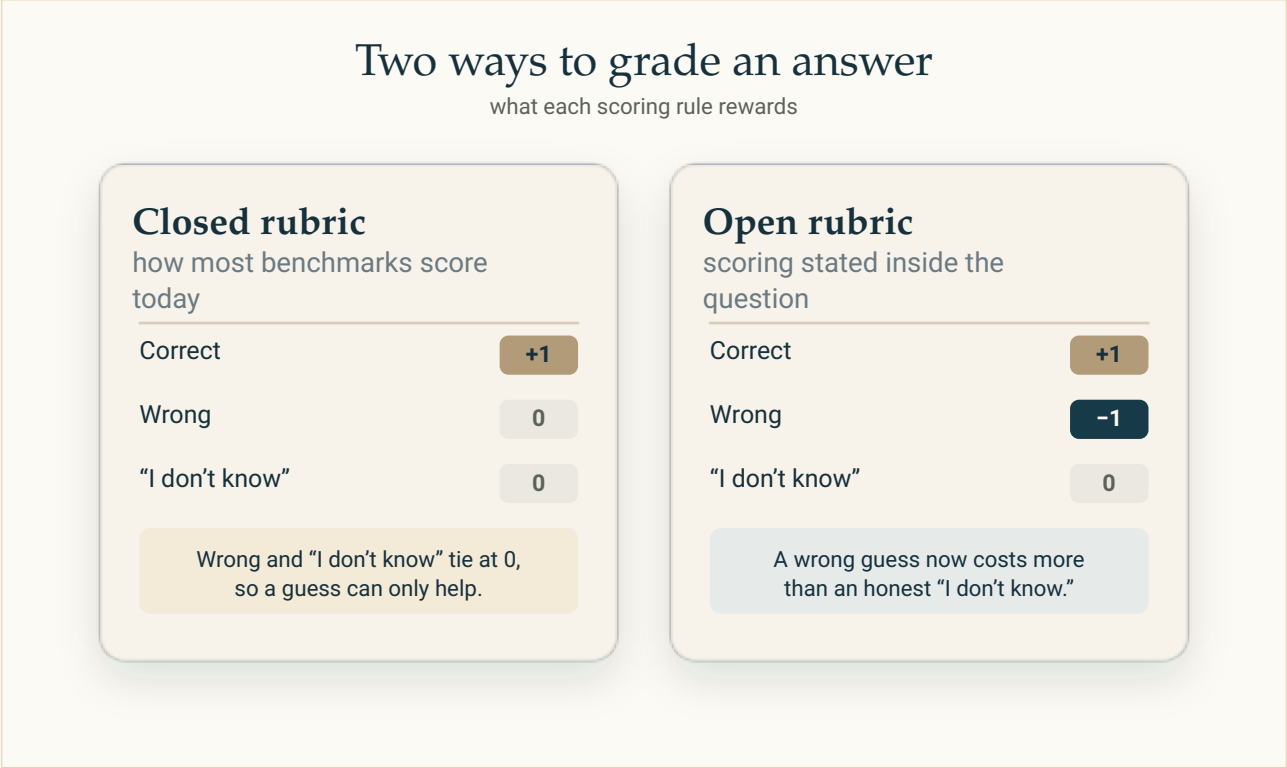
チューリングがまったく別の問題で実際に直面した問いはこうだ。次に引く球が、これまで一度も見ていない色である確率はどれくらいか？一度も引いていないものは直接数えられない。しかし、**一度だけ見た色**、つまり「シングルトン」は数えられる。**Good–Turing 推定**と呼ばれる方法では、標本中でシングルトンが占める割合が、まだ一度も観測していない色に隠れている確率質量を推定する。100 回のうち 10 回が一度きりの色だったなら、次の球がまったく新しい色である確率はおよそ **10 / 100 = 10%** となる。

一度しか見ていない色は誤りではない。それは、自分がどれだけ知らないかを測る手掛かりだ。一度だけ現れる色が多いということは、「世界にはまだ、単に引いていないものがたくさんある」と標本が教えているのである。

次に、色を誕生日へ、袋をモデルの学習テキストへ置き換える。学習中に見た誕生日のうち、**5 件に 1 件が一度しか現れなかった**としよう。同じ考え方を使えば、確率の約 5 分の 1 はモデルが事実上見たことのない誕生日にあることになる。そして誕生日には、代わりに頼れる規則性がない（人の誕生日を推論で計算することはできない）。したがって、一度しか見て

> いない日付や一度も見ていない日付については、モデルには適切に重みをつける材料がなく、およそ **5 件に 1 件**は間違える。どれほど賢くても、この問題は消せない。学ぶべき情報がなかったからだ。
> 議論全体を小さくまとめるとこうなる。シングルトン率は、このデータだけではどれだけの事実を学べないかの目安になり、それが誤り率の**下限**につながる。そして、モデルが首都をほとんど間違えない理由も同じだ。*Paris* は何度も登場し、シングルトン率はほぼゼロなので、学べる情報が十分にある。

これでハルシネーションがどこから生まれるかは説明できる。しかし、なぜそれが残り続けるのか —— 役に立ち、正直なモデルにするための後段学習を経た後も、なぜ疑いを認めずにハッタリを言うのか —— は説明できない。ここで論文が使うたとえは、少し居心地が悪いほど的確だ。試験中、答えが分からない学生を想像してほしい。空欄なら 0 点、当てずっぽうでも正解すれば 1 点なら、点数を最大化する行動は「答える」ことだ。自信ありげに具体的に書き、「分かりません」とは言わない。学生はそれを学ぶ。そしてモデルも同じことを学ぶ —— 私たちが同じように採点しているからだ。著者たちは、この分野が実際に競争に使っているベンチマーク、モデルが順位を上げるよう調整されるリーダーボードを調べ、そのほぼすべてが「分かりません」に誤答とまったく同じ点数、つまり 0 点を与えることを見いだした。そのルールでは、常に推測するモデルのほうが、不確実さを正直に示す点だけが異なるモデルより成績がよくなる。かなり文字通りの意味で、私たちはモデルをそう採点してきたのである。



ベンチマークは、推測を合理的な行動にしてしまう。多くのベンチマークで使われる採点（左）では、誤答も正直な「分かりません」も 0 点なので、推測は得にしかならない。質問文の中でルールを明示し、誤答にペナルティーを設ける（右）と、自信がないときに回答を控えるほうが有利になり得る。変わるのはテストが何を報いるかであり、それだけでハルシネーションそのものが解決するわけではない。

Original diagram —The Clean Paper CC BY 4.0

ここは覚えておく価値がある。よくある見出しとは逆を向いているからだ。ハルシネーションはしばしば、この技術に避けがたく付きまとう、ほとんど神秘的な限界として語られる。この論文はその両方に異議を唱える。事前学習で生じる誤りの下限は謎ではない。機械学習が何十年も扱ってきた、ごく普通の統計的誤差だ。そして、それが後まで残ることも必然ではない。少なくとも一部は、私たち自身が作ったインセンティブであり、変えることができる。自信がないときに単に回答を控えるシステムなら、ハルシネーションは起こさない。実運用モデルがそう振る舞わない理由の一つは、私たちの採点表が回答拒否を罰することにある。

著者たちは何をしたのか

論文は三つの部分からなる。第一に、「妥当なテキストだけを生成する」ことは、二値分類問題である「この文は妥当か？」に少なくとも同程度に難しいことを示し、事前学習中に一定のハルシネーションが統計的に避けられないという数学的議論を構築した。第二に、影響力の大きい 10 個のベンチマークを調べ、主流の正解率型指標が回答保留より推測を報いると論じた。第三に、修正案とそのケーススタディーを示した。それが**オープン・ルーブリック**評価である。採点法を質問文の中に明示する（たとえば「正解は 1 点、誤答は –1 点。確信度が 50% 未満なら回答を控えること」）ことで、正直さが報われる条件をモデル自身が分かるようにする。著者らは Google の Gemini 3 Pro、OpenAI の GPT-5、xAI の Grok 4、Anthropic の Claude Opus 4.5 という四つのフロンティアモデルに対し、SimpleQA の 4,326 問の事実問題を用いてこれを試した。なお著者らは、このケーススタディーが例示目的であり、「モデル間の統制された評価ではない」と明記している（デフォルト設定で、チューニングなし、コストの正規化なし）。

何が分かったのか

- **事前学習は一定の誤りを不可避にする**。モデルが自信を持って偽の文を出す率には、それを使って作れる最良の「この文は妥当か？」分類器の誤り率のおおむね 2 倍に対応する下限がある。学べるパターンのない事実では、その下限は少なくともシングルトン率 —— 学習中に一度しか現れない事実の割合 —— になる。データが完全にきれいでも、ある程度のハルシネーションは避けられない。
- **採点は具体的に推測を報いる**。通常の正解／不正解採点では、回答を一度も控えないことが最適戦略になる。そして著者らの調査では、多くの人気ベンチマークが「分かりません」を単なる不正解として扱っている。著者ら自身の例は鮮烈だ。SimpleQA では、生の正解率は OpenAI の o4-mini をわずかに有利にする。このモデルはほぼすべての問いに答え、4 分の 3 以上を間違える。一方 GPT-5-mini は、不確かなときに回答を控えるため誤答がはるかに少ない。それでも、無謀なモデルのほうがランキング上ではよく見える。
- **オープン・ルーブリックはインセンティブを反転させる（このケーススタディーでは）**。著者らは単純なハルシネーション抑制策 —— モデルに二回答えさせ、二つが食い違えば回答を控えさせる —— を試した。通常の正解率では、この方法は誤答を減らす一方で正解率も下げるため、指標そのものがその導入を不利にする。ところがオープン・ルーブリックでは、複数の誤答ペナ

ルティー設定にわたり四モデルすべてで同じ抑制策が有利になった。また、生の正解率では不確かなときに回答を控えるため不利だった GPT-5-mini が、採点規則を明示すると o4-mini より上になる（各モデル n = 4,326 問）。

これはおそらく何を意味するのか

ハルシネーションを減らすために必要なのは、主として新しい「ハルシネーション専用テスト」を発明することではない。主流ベンチマークが不確実性をどう採点するかを変え、「分かりません」と認めることが罰にならないようにすることだ。評価の仕組みが変わらない限り、ハルシネーションを減らすことは正解率ポイントを失う行為であり続け、そのため改善が抑制され続ける。著者らがこの問題を「社会技術的（socio-technical）」と呼ぶのはそのためだ。より良い指標を作ることと、影響力のあるランキングにそれを採用させることの両方が必要になる。

この研究からは言えないこと

- オープン・ルーブリックが実世界でハルシネーションを解決すると示したわけでは**ない**。裏付けとなる実験は、四モデルをデフォルト設定で使い、選んだ一つの抑制策を一つの事実 QA テストで試した、小規模で意図的に**非統制**のケーススタディーである。目的はインセンティブが反転することを示すことであって、モデル順位や一般的な有効性を証明することではない。
- 採点が唯一の原因だと主張しているわけでは**ない**。学習データ中の誤り、本当に難しい問題、馴染みのないプロンプトなどは別の原因として残る。
- 「ハルシネーションは不可避だ」というよくある言い方を支持しているわけでは**ない**。著者らの主張は逆で、検証可能な問いにだけ答え、それ以外は「分かりません」と言うシステムならハルシネーションを起こさない、と論じる。
- 事前学習にある誤りの下限そのものが消えるわけでは**ない**。論文はそれを説明し、上限ではなく下限を与える。また、その下限が対象とするのは自信を伴う事実誤りであり、モデルのあらゆる挙動ではない。
- オープン・ルーブリックだけで十分だと示したわけでは**ない**。変わるのは評価が何を報いるかであり、検索、ツール利用、より適切に較正されたモデルの代わりになるものではない。

証拠をどう評価するか

- 中核は**数学**であり、測定ではなく形式的な下限である。理論的議論としては、その仮定の範囲内で自立している。
- 問題を**意図的に単純化したモデル**に依存している。著者ら自身、すべての回答を「正解・不正解・分かりません」の三つに分けることを「偽の三分法（false trichotomy）」と呼び、もっともきれいな下限を導くための「任意事実」設定が理想化されていることも明示している。
- ベンチマーク調査は**小規模で選定された標本**である。影響力のある 10 評価を見たもので、網羅

的監査ではない。
- ケーススタディーは**実データだが限定的**である。四つのフロンティアモデル、一つの抑制策、SimpleQA のみ、デフォルト設定で、明示的に「統制された評価ではない」。インセンティブ議論の概念実証であって、ベンチマーク結果ではない。
- 視点の所在も明記しておく価値がある。四人の著者のうち三人は OpenAI の現職または元職員であり、論文はモデル評価方法を分野全体で変えるべきだと主張している。論証された立場ではあるが、無関係な第三者による中立レビューではない。だから退けるのではなく、その点も含めて重みづけすべきである。（公平を期すなら、論文は他社だけでなく自社の o4-mini と GPT-5-mini にも同じ批判を向けている。）

なぜ重要なのか

過度に騒がれてきた問題を、別の枠で見せるからだ。「ハルシネーション」は、不気味な欠陥か動かせない壁のどちらかとして語られがちだ。この論文はそれを、普通のものであり、一部は自分たちが作った問題だとする。理解可能な統計的下限の上に、私たちが選んだインセンティブが乗っているというわけだ。より広い教訓は、静かだが有用である。信頼性をさらに高めるには、何を作るかだけでなく、何を測るかが同じくらい重要なのかもしれない。

要点

言語モデルが自信を持って偽の答えを返す理由には二つある。第一は統計的なものだ。学べる規則性のない事実では、言語を模倣するよう学習したモデルはときに間違え、その下限は、たとえば学習中に一度しか現れない事実の割合から推定できる。第二はインセンティブだ。モデル順位に使われるほぼすべてのベンチマークは「分かりません」を誤答と同じ点数にするため、常に推測するほうが有利になる。実際、4 分の 3 以上を間違えるモデルが、不確かなときに回答を控えるより正直なモデルより高得点になることすらある。著者らの提案は別のハルシネーションテストではなく、「オープン・ルーブリック」—— 採点規則を質問内に明示することだ。四つのフロンティアモデルを使ったケーススタディーでは、それによってインセンティブが反転し、ハルシネーションを減らす方法が罰ではなく報酬を受けるようになった。これは理論＋調査に、小規模で明示的に非統制の実験を加えた *Nature* 査読済み論文である。提案は有望だが、大規模に有効とはまだ示されていない。そして著者らは、ハルシネーションは神秘的でも、厳密な意味で不可避でもないと論じている。

冷静に見ると

論文が示したこと：事前学習中に一定のハルシネーションが生じる数学的下限（規則性のない事実では少なくとも「シングルトン率」）、主要ベンチマークの多くが「分かりません」に点を与えないという調査結果、そして採点法をプロンプト内に明記する「オープン・ルーブリック」によって、通常の正解率では不利だったハルシネーション抑制策が四モデルのケーススタディーで有利になる

こと。

もっともらしいが未証明のこと：オープン・ルーブリックを主流ベンチマークに採用すれば、実運用モデルのハルシネーションも有意に減ること。裏付けとなる実験は小規模で、明示的に非統制である。

論文が示していないこと：ハルシネーションが不可避であること（むしろ逆を論じる）、採点が唯一の原因であること、ハルシネーションを完全に消せること、ケーススタディーが四モデルを相互に順位づけしていること。

主な限界：正解／不正解／「分かりません」という意図的に単純化したモデル（著者自身が「偽の三分法」と呼ぶ）、小規模で選定されたベンチマーク調査（10 評価）、非統制のケーススタディー（四モデル、一つの抑制策、一つのテスト、デフォルト設定）、そして評価方法をどう変えるべきかについて OpenAI 関係者が中心となって論じている点。

一般読者はどの程度確信してよいか？ハルシネーションは神秘的でも厳密に不可避でもなく、主流ベンチマークが現在は推測を報いる構造になっている、という点には高い確信を持ってよい。提案された修正が有効だという点は中程度。実際の概念実証はあるが、広範かつ大規模に働くことはまだ示されていない。

出典

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