The Clean Paper

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

Por que modelos de linguagem alucinam — e por que a forma como os avaliamos mantém isso assim

As falsidades confiantes que chamamos de "alucinações" não são uma falha misteriosa: algumas são um subproduto estatístico do treinamento, e persistem porque benchmarks centrais recompensam um chute confiante mais do que um "não sei" honesto. Um estudo de caso com quatro modelos de fronteira mostra que declarar as regras de pontuação no prompt ("rubricas abertas") inverte esse incentivo.

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

Por que os modelos chutam, e por que nós os ensinamos a fazer isso

Pergunte a um grande modelo de linguagem o aniversário de uma pessoa desconhecida e ele pode responder “7 de março” com a confiança estável de quem está lendo de um cartão — e estar errado, três vezes seguidas, com três datas diferentes. Os autores dão exatamente esse tipo de exemplo: modelos líderes recebem uma pergunta factual simples — o aniversário de uma pessoa, ou o significado de uma sigla obscura — e cada um inventa com confiança uma resposta diferente, nenhuma correta. A palavra da indústria para isso é *alucinação*, o que faz parecer uma falha de percepção. O primeiro movimento do artigo é tirar o mistério disso.

Comece por como um modelo é construído. Em sua primeira e maior etapa de treinamento, ele aprende, em essência, como a linguagem fluente se parece lendo uma quantidade enorme de texto. Agora pegue um fato sem padrão por trás — o aniversário de uma pessoa específica. Se essa data apareceu no texto de treinamento uma vez, ou nunca, não há nada a que um aprendiz de padrões possa se agarrar: a resposta é, do ponto de vista do modelo, arbitrária. Os autores tornam isso preciso tomando emprestada uma ideia antiga (de Alan Turing, para um problema diferente): se um em cada cinco aniversários aparece apenas uma vez nos dados, deve-se esperar que um modelo erre pelo menos um em cada cinco deles — não porque esteja quebrado, mas porque nunca havia ali algo a aprender. (Pela mesma lógica, modelos quase nunca erram a capital de um país: essas aparecem o tempo todo.) Eles argumentam, com certo cuidado, que distinguir uma afirmação verdadeira de uma falsa plausível já é em si um problema difícil, e que produzir *apenas* afirmações verdadeiras é pelo menos tão difícil quanto isso. Um piso de erro vem embutido.

A ideia por baixo: como contar o que você ainda não viu

Isto se apoia em uma ideia genuinamente engenhosa — mais antiga que os modelos de linguagem, e que vale conhecer direito.

Comece com um saco de bolas coloridas. Você não sabe quantas cores ele contém. Tira 100, uma de cada vez, e faz a contagem: vermelho 40, azul 25, verde 15, amarelo 5, roxo 3, laranja 2 — e então **dez cores diferentes que aparecem exatamente uma vez cada.**

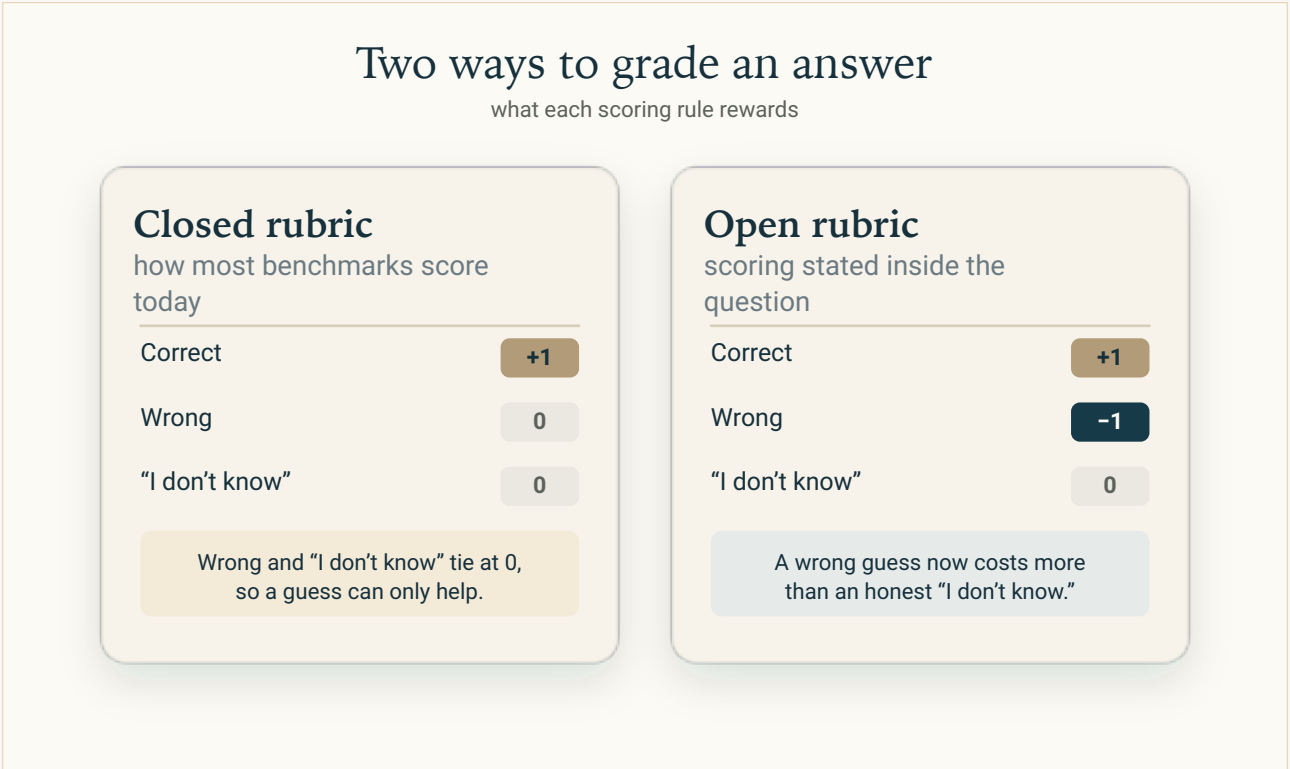
Agora a pergunta que Turing de fato enfrentou, em um problema bem diferente: *qual é a chance de a próxima bola ser de uma cor que você ainda não viu?* Você não consegue contar o que nunca sorteou — mas **consegue** contar as cores que viu exatamente uma vez, os “singletons”. O truque, chamado **estimativa de Good–Turing**, é que a fração das suas retiradas que são singletons estima a probabilidade ainda escondida nas cores que você não viu. Dez das suas cem retiradas foram cores vistas uma única vez, então a chance de a próxima bola ser de uma cor totalmente nova é cerca de **10 / 100 = 10%.**

Essas cores vistas uma só vez não são *erros*. Elas são uma medida da sua própria ignorância: muitas cores aparecendo uma vez é a maneira pela qual a amostra diz que o mundo contém mais coisas que você simplesmente ainda não sorteou.

Agora troque cores por aniversários, e o saco pelo texto de treinamento do modelo. Suponha que, entre os aniversários que ele viu, **um em cada cinco aparece exatamente uma vez.** O mesmo truque: cerca de um quinto da probabilidade vive em aniversários que o modelo, na

> prática, nunca viu — e um aniversário não tem um padrão ao qual recorrer (não dá para *calcular* o aniversário de alguém). Portanto, uma data vista uma vez, ou nunca, é uma moeda que o modelo não consegue ponderar, e ele errará aproximadamente **um em cada cinco** desses casos. Nenhuma esperteza corrige isso: não havia nada ali a aprender.
> Esse é o argumento inteiro em miniatura: a taxa de singletons mede quanto do mundo é inapreensível *a partir destes dados*, e isso se torna um **piso** para os erros. É também por isso que um modelo quase nunca erra uma capital — *Paris* aparece constantemente, sua taxa de singleton é quase zero, então há muito o que aprender.

Isso explica de onde vêm as alucinações. Não explica por que elas *sobrevivem* — por que os modelos, depois de todo o treinamento posterior feito para torná-los úteis e honestos, ainda blefam em vez de admitir dúvida. Aqui a analogia do artigo é quase desconfortavelmente apropriada. Imagine um aluno em uma prova que não sabe uma resposta. Se deixar em branco vale zero e um chute pode valer um, a jogada que maximiza a nota é chutar — com confiança, de forma específica, nunca "não tenho certeza". Alunos aprendem isso. Modelos, ao que parece, também — porque nós os avaliamos do mesmo jeito. Os autores passaram pelos benchmarks em que o campo de fato compete, os rankings que os modelos são ajustados para subir, e descobriram que quase todos dão a "não sei" exatamente a mesma pontuação de uma resposta errada: zero. Sob essa regra, um modelo que sempre chuta vence um modelo idêntico que sinaliza honestamente sua incerteza. Em um sentido bastante literal, nós os pontuamos para chegar a isso.



Benchmarks podem tornar o chute racional. Na pontuação usada pela maioria dos benchmarks (à esquerda), uma resposta errada e um "não sei" honesto valem zero — então um chute só pode ajudar. Declare as regras na pergunta, com penalidade por errar (à direita), e abster-se quando há incerteza pode se tornar a melhor jogada. Isso muda o que o teste recompensa; por si só, não resolve a alucinação.

Esta é a parte que vale guardar, porque vai contra a manchete usual. A alucinação costuma ser vendida como um limite inevitável, quase místico, da tecnologia. O artigo contesta as duas coisas. O piso do pré-treinamento não é um mistério — é erro estatístico comum, do tipo que o aprendizado de máquina entende há décadas. E a persistência não é inevitável — é, em parte, um incentivo que construímos e poderíamos mudar. Um sistema que simplesmente se recusasse a responder quando não tivesse certeza não alucinaria; a razão pela qual modelos em uso não se comportam assim é que nossos placares punem a recusa.

O que os autores fizeram

O artigo tem três partes. Primeiro, um argumento matemático de que alguma alucinação é estatisticamente forçada durante o pré-treinamento, mostrando que "gerar apenas texto válido" é pelo menos tão difícil quanto um problema binário de classificação "esta afirmação é válida?". Segundo, um argumento — apoiado por um levantamento de dez benchmarks influentes — de que métricas convencionais do tipo acerto/erro recompensam chutar em vez de abster-se. Terceiro, uma correção proposta e um estudo de caso que a testa: avaliações de **rubrica aberta**, em que a pontuação é declarada dentro da própria pergunta (por exemplo, "uma resposta correta vale 1, uma errada vale −1, então abstenha-se se tiver menos de 50% de certeza"), para que o modelo possa perceber quando a honestidade está sendo recompensada. Eles testam isso em quatro modelos de fronteira — Gemini 3 Pro, do Google; GPT-5, da OpenAI; Grok 4, da xAI; e Claude Opus 4.5, da Anthropic — usando as 4.326 perguntas factuais do SimpleQA. Eles deixam explícito que o estudo de caso é ilustrativo, "não uma avaliação controlada entre modelos" (configurações padrão, sem ajuste, sem normalização de custo).

O que encontraram

- **O pré-treinamento força algum erro.** A taxa com que um modelo emite falsidades confiantes é limitada inferiormente por (aproximadamente duas vezes) a taxa de erro do melhor classificador "esta afirmação é válida?" construído a partir dele. Para fatos sem padrão aprendível, esse piso é pelo menos a *taxa de singletons* — a fração dos fatos que aparecem exatamente uma vez no treinamento. Alguma alucinação é inevitável mesmo com dados perfeitamente limpos.
- **A avaliação recompensa o chute — concretamente.** Sob a pontuação comum de correto/incorreto, nunca se abster é a estratégia ótima, e o levantamento dos autores mostra que a vasta maioria dos benchmarks populares pontua "não sei" simplesmente como errado. Um exemplo vívido do próprio lado deles: no teste SimpleQA, a acurácia bruta favorece ligeiramente o o4-mini da OpenAI — que responde quase tudo e erra mais de três quartos das vezes — sobre o GPT-5-mini, que comete muito menos erros porque se abstém quando não tem certeza. O modelo mais imprudente parece melhor no placar.
- **Rubricas abertas invertem o incentivo (no estudo de caso deles).** Eles testam uma mitigação

simples de alucinação (fazer o modelo responder duas vezes e abster-se se as duas respostas discordarem). Sob acurácia padrão, a mitigação reduz erros *mas também reduz a acurácia* — então a métrica desestimula adotá-la. Sob rubricas abertas, a mesma mitigação sai na frente para todos os quatro modelos em uma faixa de penalidades; e o GPT-5-mini — que a acurácia bruta havia penalizado por se abster quando incerto — passa à frente do o4-mini quando a pontuação é declarada abertamente (n = 4.326 perguntas por modelo).

O que isso provavelmente significa

Reduzir alucinação, em grande parte, não é questão de inventar mais testes específicos de alucinação. É questão de mudar como os benchmarks centrais pontuam a incerteza, para que admitir "não sei" deixe de ser punido. Enquanto o placar não mudar, reduzir alucinação continuará custando pontos de acurácia aos modelos e, portanto, continuará sendo desencorajado — por isso os autores enquadram o problema como "sociotécnico": em parte métrica melhor, em parte fazer os rankings influentes adotá-la.

O que isto *não* prova

- Não mostra que rubricas abertas resolvem alucinação no uso real. O experimento de apoio é um estudo de caso pequeno, deliberadamente **não controlado** — quatro modelos em configurações padrão, uma mitigação escolhida, um teste de QA factual — destinado a demonstrar a *inversão do incentivo*, não a ranquear modelos nem provar eficácia geral.
- Não afirma que a avaliação é a *única* causa. Erros nos dados de treinamento, problemas genuinamente difíceis e prompts desconhecidos continuam sendo fontes separadas.
- Não apoia a frase popular de que alucinações são inevitáveis. Os autores argumentam o contrário: um sistema que respondesse apenas perguntas verificáveis e, caso contrário, dissesse "não sei" nunca alucinaria.
- Não faz desaparecer o piso do pré-treinamento — ele o explica e delimita, e o limite diz respeito a erros factuais confiantes, não a todo comportamento do modelo.
- Não mostra que rubricas abertas sejam *suficientes* por si só. Elas mudam o que uma avaliação recompensa; não substituem recuperação, uso de ferramentas ou modelos mais bem calibrados.

Quão forte é a evidência?

- O núcleo é **matemática** — limites inferiores formais, não medições. Como argumento teórico, é sólido em seus próprios termos.
- Ele se apoia em **modelos deliberadamente simplificados** do problema; os próprios autores destacam a "falsa tricotomia" de tratar toda resposta como correta, incorreta ou "não sei", e o cenário idealizado de "fatos arbitrários" usado para o limite mais limpo.

- O levantamento de benchmarks é uma **amostra pequena e curada** — dez avaliações influentes, não uma auditoria exaustiva.
- O estudo de caso é **real, mas limitado**: quatro modelos de fronteira, uma única mitigação, apenas SimpleQA, configurações padrão, explicitamente "não uma avaliação controlada". É uma prova de conceito para o argumento de incentivo, não um resultado de benchmark.
- Vale nomear o ponto de vista: três dos quatro autores são ou foram empregados da **OpenAI**, e o artigo argumenta que o campo deve mudar a forma como avalia modelos. É uma posição bem argumentada de uma parte interessada, não uma revisão externa neutra — algo a pesar, não a descartar. (Para crédito do artigo, a crítica mira seus próprios modelos, o4-mini e GPT-5-mini, tão prontamente quanto os outros.)

Por que isso importa

Ele reenquadra um problema fortemente inflado. "Alucinação" tende a ser vendida ou como um defeito fantasmagórico, ou como um muro imóvel; este artigo a torna comum e em parte autoinfligida — um piso estatístico que podemos de fato entender, sentado sobre um incentivo que escolhemos. A lição mais ampla é mais quieta e mais útil: o progresso em confiabilidade pode depender tanto de *o que medimos* quanto do que construímos.

Resumo claro

Respostas falsas confiantes de modelos de linguagem vêm de dois lugares. O primeiro é estatístico: quando um fato não tem padrão a aprender, um modelo treinado para imitar linguagem às vezes o errará, e esse piso pode ser estimado (por exemplo, a partir de quantos fatos aparecem apenas uma vez no treinamento). O segundo são incentivos: quase todo benchmark em que modelos são ranqueados pontua "não sei" igual a uma resposta errada, então chutar sempre vence — a ponto de um modelo errado em três quartos das vezes poder superar outro mais honesto que se abstém. A proposta dos autores não é mais um teste de alucinação, mas "rubricas abertas": declarar a pontuação dentro da pergunta. Em um estudo de caso com quatro modelos de fronteira, isso inverte o incentivo, de modo que um método que reduz alucinação é recompensado em vez de penalizado. É um artigo de teoria mais levantamento, com um experimento pequeno e explicitamente não controlado, revisado por pares na *Nature*; a correção é promissora, mas ainda não demonstrou funcionar em escala, e as alucinações são apresentadas como nem misteriosas nem estritamente inevitáveis.

Checagem sem enrolação

O que o artigo mostra: Um limite inferior matemático pelo qual alguma alucinação é forçada durante o pré-treinamento (pelo menos a "taxa de singletons" para fatos sem padrão); um levantamento mostrando que a maioria dos benchmarks líderes não dá crédito a "não sei"; e um estudo de caso

com quatro modelos em que declarar a pontuação no prompt (“rubricas abertas”) faz vencer um método de redução de alucinação, onde a acurácia simples o havia penalizado.

O que é plausível, mas não provado: Que rubricas abertas, incorporadas aos benchmarks centrais, reduziriam de modo significativo a alucinação em modelos implantados. O experimento de apoio é pequeno e explicitamente não controlado.

O que ele não mostra: Que alucinações são inevitáveis (ele argumenta o inverso); que a avaliação é a única causa; que a alucinação pode ser eliminada de uma vez; que o estudo de caso ranqueia os quatro modelos entre si.

Principais limitações: Um modelo deliberadamente simplificado correto/incorreto/“não sei” (os autores o chamam de “falsa tricotomia”); um levantamento pequeno e curado de benchmarks (dez avaliações); um estudo de caso não controlado (quatro modelos, uma mitigação, um teste, configurações padrão); e um argumento liderado pela OpenAI sobre como o campo deveria avaliar modelos.

Quanta confiança um leitor geral deve ter? Alta de que a alucinação não é misteriosa nem estritamente inevitável, e de que benchmarks centrais hoje recompensam o chute. Moderada de que a correção proposta ajude — agora ela tem uma prova de conceito real, mas ainda não uma demonstração de que funciona de forma ampla e em escala.

Fontes

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