The Clean Paper

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

Porque é que os modelos de linguagem alucinam — e porque é que a forma como os avaliamos mantém isso assim

As falsidades confiantes que chamamos "alucinações" não são uma falha misteriosa: algumas são um subproduto estatístico do treino, e persistem porque benchmarks centrais recompensam um palpite 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 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

Porque é que os modelos dão palpites, e porque é que nós os ensinamos a fazê-lo

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 a ler 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 isto é *alucinação*, o que faz parecer uma falha de perceção. O primeiro movimento do artigo é tirar o mistério disso.

Comece por como um modelo é construído. Na sua primeira e maior etapa de treino, ele aprende, em essência, como se parece a linguagem fluente lendo uma quantidade enorme de texto. Agora pegue num facto sem padrão por trás — o aniversário de uma pessoa específica. Se essa data apareceu no texto de treino uma vez, ou nunca, não há nada a que um aprendiz de padrões possa agarrar-se: a resposta é, do ponto de vista do modelo, arbitrária. Os autores tornam isto 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 esperar-se que um modelo erre pelo menos um em cada cinco deles — não porque esteja avariado, mas porque nunca havia ali algo a aprender. (Pela mesma lógica, modelos quase nunca erram a capital de um país: essas aparecem a toda a hora.) Eles argumentam, com algum 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 como isso. Um piso de erro vem incorporado.

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

Isto assenta numa ideia genuinamente engenhosa — mais antiga do que os modelos de linguagem, e que vale a pena conhecer bem.

Comece com um saco de bolas coloridas. 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 depois **dez cores diferentes que aparecem exatamente uma vez cada.**

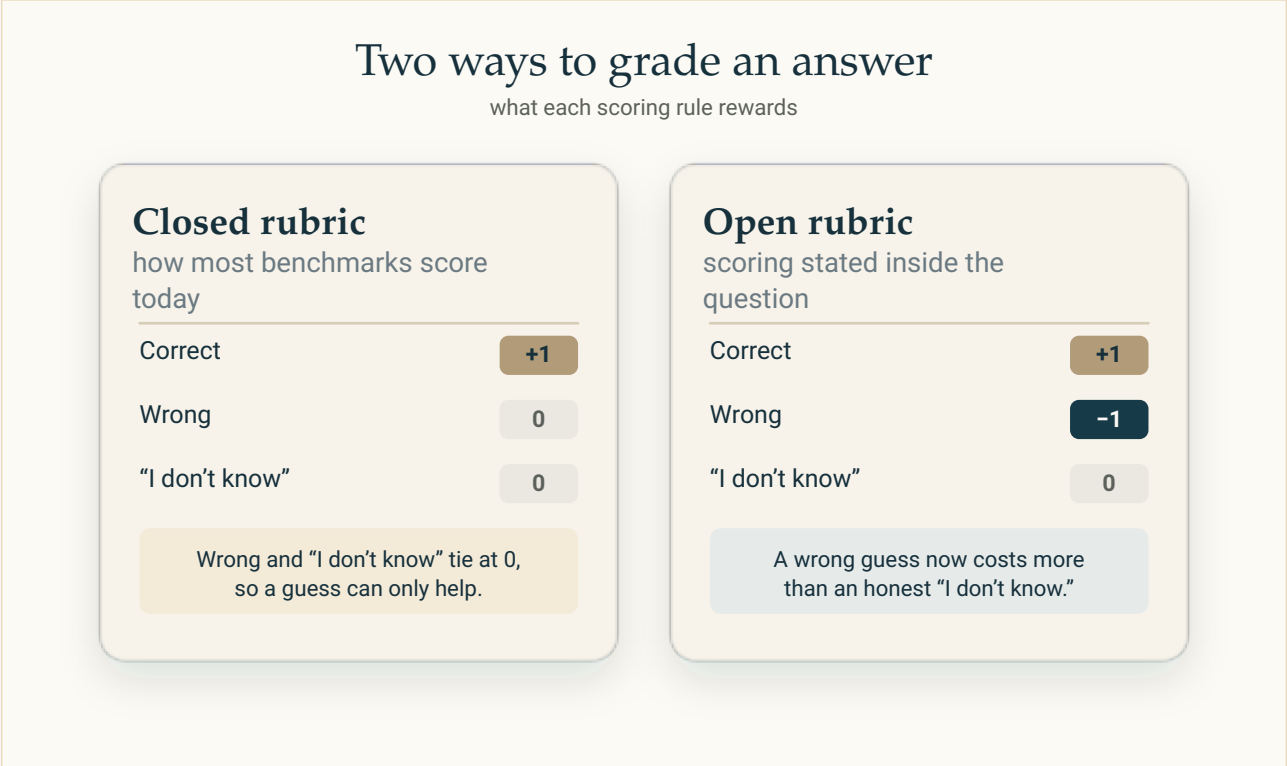
Agora a pergunta que Turing de facto enfrentou, num problema bem diferente: *qual é a probabilidade de a próxima bola ser de uma cor que ainda não viu?* Não consegue contar o que nunca sorteou — mas **consegue** contar as cores que viu exatamente uma vez, os "singletons". O truque, chamado **estimativa de Good–Turing**, é que a fração das suas retiradas que são singletons estima a probabilidade ainda escondida nas cores que não viu. Dez das suas cem retiradas foram cores vistas uma única vez, logo a probabilidade 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*. São uma medida da sua própria ignorância: muitas cores a aparecer uma vez é a forma como a amostra diz que o mundo contém mais coisas que simplesmente ainda não sorteou.

Agora troque cores por aniversários, e o saco pelo texto de treino 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 a que 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 isto: 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 torna-se um **piso** para os erros. É também por isso que um modelo quase nunca erra uma capital — *Paris* aparece constantemente, a sua taxa de singleton é quase zero, logo há muito o que aprender.

Isto explica de onde vêm as alucinações. Não explica porque é que elas *sobrevivem* — porque é que os modelos, depois de todo o treino posterior feito para os tornar úteis e honestos, ainda fazem bluff em vez de admitir dúvida. Aqui a analogia do artigo é quase desconfortavelmente apropriada. Imagine um aluno num exame que não sabe uma resposta. Se deixar em branco vale zero e um palpite pode valer um, a jogada que maximiza a nota é arriscar — com confiança, de forma específica, nunca "não tenho a certeza". Alunos aprendem isto. Modelos, ao que parece, também — porque nós os avaliamos da mesma forma. Os autores passaram pelos benchmarks em que o campo de facto 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 esta regra, um modelo que arrisca sempre vence um modelo idêntico que sinaliza honestamente a sua incerteza. Num sentido bastante literal, pontuámo-los para chegar a isto.



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

Esta é a parte que vale a pena guardar, porque vai contra a manchete habitual. A alucinação costuma ser vendida como um limite inevitável, quase místico, da tecnologia. O artigo contesta as duas coisas. O piso do pré-treino não é um mistério — é erro estatístico comum, do tipo que a aprendizagem automática 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 a certeza não alucinaria; a razão pela qual modelos em uso não se comportam assim é que as nossas tabelas de pontuação 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-treino, mostrando que "gerar apenas texto válido" é pelo menos tão difícil como 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 certo/errado recompensam arriscar 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, portanto abstenha-se se tiver menos de 50% de certeza"), para que o modelo possa perceber quando a honestidade está a ser recompensada. Eles testam isto em quatro modelos de fronteira — Gemini 3 Pro, da 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-treino 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 factos sem padrão aprendível, esse piso é pelo menos a *taxa de singletons* — a fração dos factos que aparecem exatamente uma vez no treino. Alguma alucinação é inevitável mesmo com dados perfeitamente limpos.
- **A avaliação recompensa o palpite — 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 exatidão bruta favorece ligeiramente o o4-mini da OpenAI — que responde quase tudo e erra mais de três quartos das vezes — em relação ao GPT-5-mini, que comete muito menos erros porque se abstém quando não tem a certeza. O modelo mais imprudente parece melhor na tabela.
- **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 exatidão padrão, a mitigação reduz erros *mas também reduz a exatidão* — portanto a métrica desincentiva adotá-la. Sob rubricas abertas, a mesma mitigação sai à frente para todos os quatro modelos numa faixa de penalizações; e o GPT-5-mini — que a exatidão 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 isto 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 a tabela de pontuação não mudar, reduzir alucinação continuará a custar pontos de exatidão aos modelos e, portanto, continuará a ser 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 classificar modelos nem a provar eficácia geral.
- Não afirma que a avaliação é a *única* causa. Erros nos dados de treino, problemas genuinamente difíceis e prompts desconhecidos continuam a ser fontes separadas.
- Não apoia a frase popular de que alucinações são inevitáveis. Os autores argumentam o contrário: um sistema que respondesse apenas a perguntas verificáveis e, caso contrário, dissesse "não sei" nunca alucinaria.
- Não faz desaparecer o piso do pré-treino — explica-o e delimita-o, e o limite diz respeito a erros factuais confiantes, não a todo o 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 nos seus próprios termos.
- Apoia-se 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 "factos 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 a pena 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 os seus próprios modelos, o4-mini e GPT-5-mini, tão prontamente quanto os outros.)

Porque é que isto importa

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

Resumo claro

Respostas falsas confiantes de modelos de linguagem vêm de dois lugares. O primeiro é estatístico: quando um facto não tem padrão a aprender, um modelo treinado para imitar linguagem por vezes irá errá-lo, e esse piso pode ser estimado (por exemplo, a partir de quantos factos aparecem apenas uma vez no treino). O segundo são incentivos: quase todo benchmark em que modelos são classificados pontua "não sei" igual a uma resposta errada, portanto arriscar sempre vence — ao 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. Num estudo de caso com quatro modelos de fronteira, isto 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, revisto 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-treino (pelo menos a "taxa de singletons" para factos 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 exatidão 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 em produção. 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 classifica os quatro modelos entre si.

Principais limitações: Um modelo deliberadamente simplificado correto/incorreto/"não sei" (os autores chamam-no 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 palpite. 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