The Clean Paper

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

Language models niyə hallucinate edir — və onları qiymətləndirmə üsulumuz niyə bunu davam etdirir

"Hallucinations" dediyimiz confident falsehoods sirli glitch deyil: bəziləri training-in statistik əlavə məhsuludur və mainstream benchmarks confident guess-i dürüst "I don't know"dan üstün tutduğu üçün davam edir. Dörd frontier model üzrə case study scoring rules-u prompt-un içində yazmağın ("open rubrics") həmin incentive-i çevirdiyini göstərir.

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

Modellər niyə təxmin edir — və biz onlara bunu niyə öyrətdik

Large language model-dən tanımadığı birinin ad gününü soruşun, kartdan oxuyurmuş kimi arxayın səslə “7 mart” deyə bilər — və ardıcıl üç dəfə, üç fərqli tarixlə səhv edə bilər. Müəlliflər məhz belə nümunə verir: aparıcı modellərə sadə fakt sualı — bir insanın ad günü və ya az tanınan acronym-un açılışı — verilir, hər biri əminliklə fərqli cavab uydurur və heç biri düz deyil. Sənayedə buna *hallucination* deyilir; söz sanki perception fault-dan danışır. Paper-in ilk hərəkəti sirri aradan qaldırmaqdır.

Modelin necə qurulduğundan başlayaq. İlk və ən böyük training stage-də nəhəng mətn həcmi oxuyaraq, mahiyyətcə, fluent language-in necə göründüyünü öyrənir. İndi arxasında heç bir pattern olmayan fakt götürün — konkret bir insanın ad günü. Həmin tarix training text-də bir dəfə görünibsə və ya heç görünməyibsə, pattern learner-in yapışacağı heç nə yoxdur: model baxımından cavab arbitrary-dir. Müəlliflər bunu başqa problem üçün Alan Turing-dən gələn köhnə ideya ilə dəqiqləşdirirlər: ad günlərinin beşdə biri data-da yalnız bir dəfə görünürsə, modelin onların ən azı beşdə birini səhv etməsi gözlənilməlidir — broken olduğuna görə yox, öyrənəcək heç nə olmadığına görə. (Eyni məntiqlə modellər ölkə paytaxtlarını demək olar heç vaxt səhv etmir: onlar data-da daim görünür.) Müəlliflər diqqətlə arqument edir ki, doğru ifadəni plausible false ifadədən ayırmaq özü çətin problemdir və *yalnız* doğru ifadələr yaratmaq ən azı o qədər çətindir. Error floor sistemin içinə yazılıb.

Altındakı fikir: hələ görmədiyiniz şeyi necə saymaq olar

Bu, language models-dan daha qədim, həqiqətən ağıllı bir ideyaya söykənir və onu düzgün tanımağa dəyər.

Rəngli toplar dolu torba ilə başlayın. İçində neçə rəng olduğunu bilmirsiniz. 100 top çəkirsiniz və sayırsınız: qırmızı 40, mavi 25, yaşıl 15, sarı 5, bənövşəyi 3, narıncı 2 — sonra isə **hərəsi yalnız bir dəfə çıxan on fərqli rəng**.

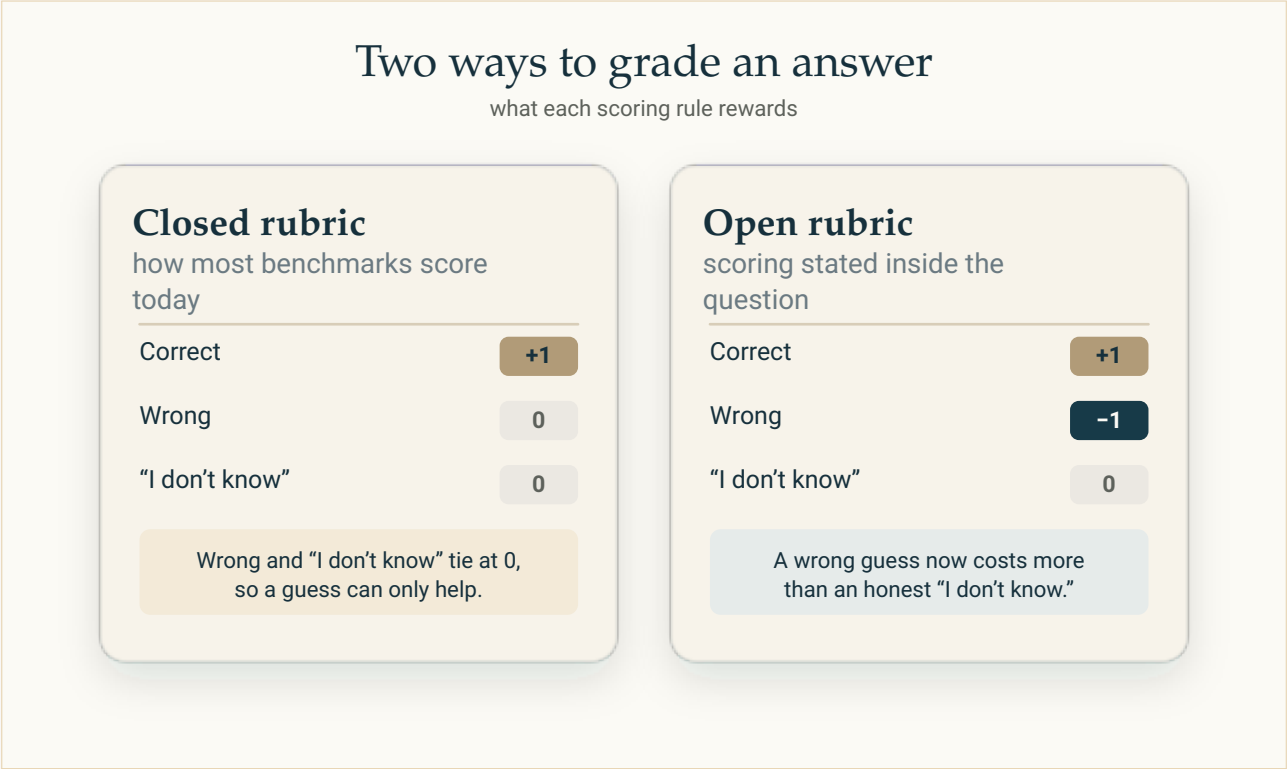
Turing-in tamam başqa problemdə qarşılaşdığı sual: *növbəti topun indiyə qədər heç görmədiyiniz rəngdə olma ehtimalı nədir?* Heç çəkmədiyinizi birbaşa saya bilməzsiniz — amma **yalnız bir dəfə gördüyünüz rəngləri**, “singletons”u saya bilərsiniz. **Good–Turing estimation** adlanan hiylə budur: seçimlərinizin singleton olan payı hələ görmədiyiniz rənglərdə gizlənən ehtimal kütləsini təxmin edir. 100 çəkilişdən 10-u bir dəfə görünən rəng idi, deməli növbəti topun tam yeni rəng olması ehtimalı təxminən **10 / 100 = 10%**-dir.

Bir dəfə görünən rənglər *səhv* deyil. Onlar öz məlumatsızlığınızın ölçüsüdür: bir çox rəng yalnız bir dəfə çıxırsa, sample sizə dünyada hələ çəkmədiyiniz daha çox şey olduğunu deyir.

İndi rəngləri ad günləri ilə, torbanı modelin training text-i ilə əvəz edin. Tutaq ki, modelin gördüyü ad günlərinin **beşdə biri yalnız bir dəfə görünür**. Eyni hiylə: ehtimalın təxminən beşdə biri modelin faktiki olaraq heç görmədiyi ad günlərindədir — və ad günü üçün fallback pattern yoxdur (kiminsə ad gününü *hesablaya* bilməzsiniz). Deməli bir dəfə görünən və ya heç görünməyən tarix model üçün düzgün weight verə

> bilmədiyi sikkə kimidir və o, onların təxminən **beşdə birində** səhv olacaq. Heç bir əlavə ağıllılıq bunu düzəltmir: öyrənəcək data yox idi.
> Bütün arqumentin miniaturası budur: singleton rate dünyanın *bu data-dan* öyrənilə bilməyən hissəsini ölçür və bu, errors altında **floor** olur. Buna görə model paytaxtı da demək olar səhv etmir — *Paris* davamlı görünür, singleton rate demək olar sıfırdır və öyrənəcək çox şey var.

Bu, hallucinations-ın haradan gəldiyini izah edir. Amma niyə *sağ qaldığını* — helpful və honest etmək üçün bütün sonrakı training-dən sonra modellərin niyə yenə "bilmirəm" demək əvəzinə bluff etdiyini — izah etmir. Burada paper-in analogiyası narahatedici dərəcədə uyğundur. İmtahanda cavabı bilməyən tələbəni təsəvvür edin. Boş cavab sıfır, təxmin isə bəlkə bir xal verirsə, grade-maximising davranış təxmin etməkdir — əminliklə, konkret şəkildə, heç vaxt "əmin deyiləm" demədən. Tələbələr bunu öyrənir. Məlum olur, modellər də — çünki biz onları eyni cür qiymətləndiririk. Müəlliflər sahənin həqiqətən yarışdığı benchmarks-a, modellərin qalxmaq üçün tune edildiyi leaderboards-a baxdılar və demək olar hamısının "I don't know" cavabına yanlış cavabla eyni xal — sıfır — verdiyini gördülər. Bu qayda altında həmişə təxmin edən model, qalan hər şey eyni olub uncertainty-ni dürüst göstərən modeldən üstün olacaq. Kifayət qədər literal mənada biz onları xal sistemi ilə bu davranışa itələyirik.



Benchmarks təxmin etməyi rasional edə bilər. Əksər benchmark-ların scoring-ində (solda) yanlış cavab da, dürüst "I don't know" da sıfırdır — deməli təxmin yalnız kömək edə bilər. Qaydaları sualın içində yazıb yanlış cavaba penalty verəndə (sağda), uncertainty zamanı abstain etmək daha yaxşı seçim ola bilər. Bu, testin nəyi mükafatlandırdığını dəyişir; öz-özlüyündə hallucination problemini həll etmir.

Original diagram — The Clean Paper CC BY 4.0

Yadda saxlanmalı hissə budur, çünki adi headline-a qarşı gedir. Hallucination çox vaxt texnologiyanın qaçılmaz, az qala mistik həddi kimi təqdim olunur. Paper hər iki hissəyə etiraz edir. Pretraining floor sirr deyil — machine learning-in onilliklərdir başa düşdüyü adi statistical error-dur. Davam etməsi də qaçılmaz deyil — qismən bizim qurduğumuz və dəyişə biləcəyimiz incentive-dir. Uncertain olanda sadəcə cavab verməyən sistem ümumiyyətlə hallucinate etməzdi; deployed modellərin belə davranmamasının səbəbi scoreboards-un refusal-u cəzalandırmasıdır.

Müəlliflər nə etdilər

Paper üç hissədən ibarətdir. Birincisi, pretraining zamanı bəzi hallucination-ın statistik olaraq məcburi olduğunu göstərən riyazi arqumentdir: "yalnız valid text yaratmaq" probleminin binary "bu statement valid-dirmi?" classification problemindən ən azı o qədər çətin olduğunu göstərirlər. İkincisi, on influential benchmark sorğusu ilə dəstəklənən arqument: mainstream accuracy-style metrics guessing-i abstaining-dən üstün tutur. Üçüncüsü, təklif olunan düzəliş və onu test edən case study-dir: **open-rubric** evaluations, scoring-in sualın özündə yazıldığı format (məsələn, "düz cavab 1 xal, yanlış −1, ona görə 50%-dən az əminsənsə abstain et"), beləliklə model honesty-nin nə vaxt mükafatlandırıldığını bilir. Bunu dörd frontier model-də — Google Gemini 3 Pro, OpenAI GPT-5, xAI Grok 4 və Anthropic Claude Opus 4.5 — SimpleQA-nın 4,326 fakt sualı ilə sınayırlar. Case study-nin illustrative olduğunu, "not a controlled evaluation across models" olduğunu açıq deyirlər (default settings, tuning yoxdur, cost normalisation yoxdur).

Nə tapdılar

- **Pretraining bəzi error-u məcbur edir.** Modelin əminliklə falsehood yaratma tezliyinin aşağı sərhədi ondan qurulan ən yaxşı "bu statement valid-dirmi?" classifier-in error rate-i ilə bağlıdır (təxminən ikiqat). Learnable pattern-i olmayan facts üçün bu floor ən azı *singleton rate*-dir — training-də yalnız bir dəfə görünən facts fraction. Data tam təmiz olsa belə bəzi hallucination qaçılmazdır.
- **Scoring konkret şəkildə guessing-i mükafatlandırır.** Adi correct/incorrect scoring altında heç vaxt abstain etməmək optimal strategy-dir və müəlliflərin survey-i popular benchmarks-un böyük əksəriyyətinin "I don't know"u sadəcə yanlış saydığını göstərir. Öz tərəflərindən parlaq nümunə: SimpleQA testində raw accuracy demək olar hər şeyə cavab verən və dörddə üçdən çoxunda səhv edən OpenAI o4-mini-ni uncertainty zamanı abstain etdiyi üçün xeyli az səhv edən GPT-5-mini-dən bir qədər *üstün* göstərir. Daha reckless model scoreboard-da daha yaxşı görünür.
- **Open rubrics incentive-i çevirir (case study daxilində).** Sadə hallucination mitigation sınayırlar: modelə iki dəfə cavab verdir, iki cavab fərqlidirsə abstain et. Standard accuracy altında mitigation errors-u azaldır, *amma accuracy-ni də azaldır* — deməli metric onu tətbiq etməyə qarşı incentive yaradır. Open rubrics altında eyni mitigation müxtəlif penalties

üzrə bütün dörd modeldə üstün çıxır; raw accuracy-nin unsure olanda abstain etdiyinə görə cəzalandırdığı GPT-5-mini də scoring açıq yazıldıqda o4-mini-ni keçir (hər model üçün n = 4,326 sual).

Bu, yəqin ki, nə deməkdir

Hallucination-ı azaltmaq əsasən daha çox hallucination-specific test icad etmək məsələsi deyil. Mainstream benchmarks uncertainty-ni necə score etdiyini dəyişməlidir ki, "I don't know" demək artıq cəza olmasın. Scoreboard dəyişənə qədər hallucination azaltmaq modelə accuracy points bahasına gələcək və buna görə discouraged olacaq — müəlliflərin problemi "socio-technical" adlandırmasının səbəbi də budur: bir hissə daha yaxşı metric, bir hissə influential leaderboards-un onu qəbul etməsi.

Bu nəyi *sübut etmir*

- Open rubrics-in real istifadədə hallucination-ı düzəltdiyini **göstərmir**. Dəstəkləyən experiment kiçik, qəsdən **uncontrolled** case study-dir — default settings-də dörd model, bir seçilmiş mitigation, bir factual-QA test — məqsəd *incentive flip*-i göstərməkdir, models rank etmək və ya general efficacy sübut etmək yox.
- Scoring-in *yeganə* səbəb olduğunu **iddia etmir**. Training data-dakı səhvlər, həqiqətən çətin problemlər və unfamiliar prompts ayrıca mənbələr olaraq qalır.
- "Hallucinations qaçılmazdır" kimi popular xətti **dəstəkləmir**. Müəlliflər əksini deyir: yalnız yoxlana bilən suallara cavab verən və qalanında "I don't know" deyən sistem heç vaxt hallucinate etməzdi.
- Pretraining floor-u **yox etmir** — onu izah edib bound edir; bound bütün model behaviour yox, confident factual errors haqqındadır.
- Open rubrics-in təkbaşına *kifayət etdiyini* **göstərmir**. Onlar evaluation-ın nəyi reward etdiyini dəyişir; retrieval, tool use və daha yaxşı calibrated models-in əvəzi deyil.

Dəlil nə qədər güclüdür?

- Nüvə **riyaziyyatdır** — measurements yox, formal lower bounds. Theoretical argument öz şərtləri daxilində möhkəmdir.
- Problem barədə **qəsdən sadələşdirilmiş modellərə** söykənir; müəlliflər hər response-u correct, incorrect və ya "I don't know" kimi üç yerə bölməyin "false trichotomy" olduğunu, ən təmiz bound üçün istifadə edilən "arbitrary facts" setting-in idealized olduğunu özləri qeyd edir.
- Benchmark survey **kiçik, curated sample**-dır — on influential evaluation, exhaustive audit deyil.

- Case study **realdır, amma məhduddur**: dörd frontier model, tək mitigation, yalnız SimpleQA, default settings, açıq şəkildə "not a controlled evaluation". Incentive argument üçün proof of concept-dir, benchmark result deyil.
- Vantage point-i də adlandırmaq lazımdır: dörd müəllifdən üçü **OpenAI**-da işləyir və ya işləyib və paper sahənin modelləri necə qiymətləndirdiyini dəyişməli olduğunu arqument edir. Bu, interested party-dən yaxşı əsaslandırılmış mövqedir, neutral outside review yox — nəzərə alınmalı, amma dismiss edilməməli. (Paper-in xeyrinə, tənqidi başqaları qədər öz o4-mini və GPT-5-mini modellərinə də yönəldir.)

Niyə vacibdir

Çox hyped problemi yenidən çərçivələyir. "Hallucination" ya spooky defect, ya da tərpənməz divar kimi satılır; paper onu adi və qismən özümüz yaratdığımız problemə çevirir — başa düşə bildiyimiz statistical floor və onun üstündə özümüz seçdiyimiz incentive. Daha geniş dərs daha sakit və daha faydalıdır: reliability-də sonrakı irəliləyiş nə qurduğumuz qədər *nəyi ölçdüyümüzdən* də asılı ola bilər.

Qısa xülasə

Language models-in confident false answers-i iki yerdən gəlir. Birincisi statistikdir: fact-da öyrəniləcək pattern yoxdursa, dili imitasiya etmək üçün trained model bəzən səhv edəcək və bu floor-u estimate etmək olar (məsələn, training-də yalnız bir dəfə görünən facts sayı ilə). İkincisi incentive-dir: modellərin rank olunduğu demək olar bütün benchmarks "I don't know"u yanlış cavabla eyni score edir, ona görə guessing həmişə üstün olur — o dərəcədə ki, dörddə üçdən çox səhv edən model abstain edən daha dürüst modeldən yüksək score ala bilər. Müəlliflərin təklifi yeni hallucination test yox, "open rubrics"dir: scoring-i sualın içində yazın. Dörd frontier model üzrə case study-də bu incentive-i çevirir və hallucination-reducing method cəzalandırılmaq əvəzinə reward edilir. Bu, theory-plus-survey paper-dir, kiçik və açıq şəkildə uncontrolled experiment-lə, *Nature*-da peer-reviewed; fix promising-dir, amma hələ scale-da işlədiyi göstərilməyib və paper-in arqumentinə görə hallucinations nə mistik, nə də strict mənada qaçılmazdır.

No-BS yoxlaması

Paper-in göstərdiyi: Pretraining zamanı bəzi hallucination-ı məcbur edən riyazi lower bound (pattern-siz facts üçün ən azı "singleton rate"); aparıcı benchmarks-un çoxunun "I don't know"a xal vermədiyini göstərən survey; scoring-i prompt-da yazmağın ("open rubrics") plain accuracy-nin cəzalandırdığı hallucination-reducing method-u dörd-model case study-də üstün etdiyini.

Mümkün, amma sübut olunmayan: Open rubrics mainstream benchmarks-a daxil edilsə deployed models-də hallucination-ı nəzərəçarpacaq azaldar. Dəstəkləyən experiment kiçik və açıq şəkildə uncontrolled-dur.

Göstərmədiyi: Hallucinations-ın qaçılmaz olduğunu (əksini arqument edir); grading-in yeganə səbəb olduğunu; hallucination-ın tam aradan qaldırıla biləcəyini; case study-nin dörd modeli bir-birinə qarşı rank etdiyini.

Əsas məhdudiyyətlər: Qəsdən sadələşdirilmiş correct/incorrect/"I don't know" model (müəlliflər "false trichotomy" deyir); kiçik curated benchmark survey (on evaluation); uncontrolled case study (dörd model, bir mitigation, bir test, default settings); və sahənin modelləri necə qiymətləndirməli olduğunu müdafiə edən OpenAI-led argument.

Ümumi oxucu nə qədər əmin ola bilər? Hallucination-ın nə mistik, nə strict qaçılmaz olduğuna və mainstream benchmarks-un hazırda guessing-i reward etdiyinə yüksək. Təklif olunan fix-in kömək etdiyinə orta — real proof-of-concept var, amma broad və scale nümayişi hələ yoxdur.

Mənbələr

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