The Clean Paper

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

Dil modelleri neden halüsinasyon görüyor — ve onları puanlama biçimimiz neden bunu sürdürüyor?

"Halüsinasyon" dediğimiz kendinden emin yanlışlar gizemli bir arıza değil: bazıları eğitimin istatistiksel bir yan ürünü ve kalıcı olmalarının bir nedeni de ana akım benchmark'ların dürüst bir "bilmiyorum" yerine kendinden emin tahmini ödüllendirmesi. Dört frontier model üzerinde yapılan bir vaka çalışması, puanlama kurallarını istemin içinde açıkça belirtmenin ("open rubrics") bu teşviki tersine çevirdiğini gösteriyor.

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

Modeller neden tahmin yürütür — ve biz onları neden buna alıştırdık?

Bir büyük dil modeline hiç tanımadığı birinin doğum gününü sorun; karttan okuyormuşçasına sakin bir güvenle "7 Mart" diyebilir — ve üç denemenin üçünde de, üç farklı tarihle, yanılmış olabilir. Yazarlar tam da buna benzer örnekler veriyor: önde gelen modellere basit bir olgusal soru — bir kişinin doğum günü ya da pek bilinmeyen bir kısaltmanın açılımı — sorulduğunda, her biri kendinden emin biçimde farklı bir cevap uyduruyor ve hiçbiri doğru çıkmıyor. Sektör buna *hallucination* (halüsinasyon) diyor; bu da sanki algıda bir arıza varmış izlenimi yaratıyor. Makalenin ilk yaptığı şey, bu gizem havasını ortadan kaldırmak.

Önce modelin nasıl kurulduğuna bakalım. İlk ve en büyük eğitim aşamasında model, devasa miktarda metin okuyarak akıcı dilin nasıl göründüğünü öğrenir. Şimdi arkasında hiçbir örüntü olmayan bir olgu düşünün — tek bir kişinin doğum tarihi gibi. Bu tarih eğitim metninde yalnızca bir kez geçtiyse ya da hiç geçmediyse, örüntü öğrenen bir sistemin tutunacağı hiçbir şey yoktur: model açısından cevap keyfîdir. Yazarlar bunu eski bir fikri — Alan Turing'in bambaşka bir problem için kullandığı bir fikri — ödünç alarak kesinleştiriyor: eğer doğum tarihlerinin beşte biri veride yalnızca bir kez görünüyorsa, modelin bunların en az beşte birini yanlış yapması beklenmelidir — bozuk olduğu için değil, öğrenilecek bir şey hiçbir zaman olmadığı için. (Aynı mantıkla modeller bir ülkenin başkentini neredeyse hiç yanlış söylemez: bu olgular metinde sürekli tekrar eder.) Yazarlar, doğru bir ifadeyi inandırıcı ama yanlış bir ifadeden ayırmanın başlı başına zor bir problem olduğunu ve *yalnızca* doğru ifadeler üretmenin de en az bunun kadar zor olduğunu dikkatle savunuyor. Belirli bir hata tabanı sistemin içine gömülüdür.

Altındaki fikir: henüz görmediğiniz şeyi nasıl sayarsınız?

Bu argüman, dil modellerinden çok daha eski ve gerçekten zekice bir fikre dayanıyor; düzgün biçimde tanımaya değer.

İçinde farklı renklerde toplar olan bir torbayla başlayın. Kaç renk olduğunu bilmiyorsunuz. Birer birer 100 top çekip sayıyorsunuz: kırmızı 40, mavi 25, yeşil 15, sarı 5, mor 3, turuncu 2 — sonra da **her biri yalnızca bir kez çıkan on farklı renk**.

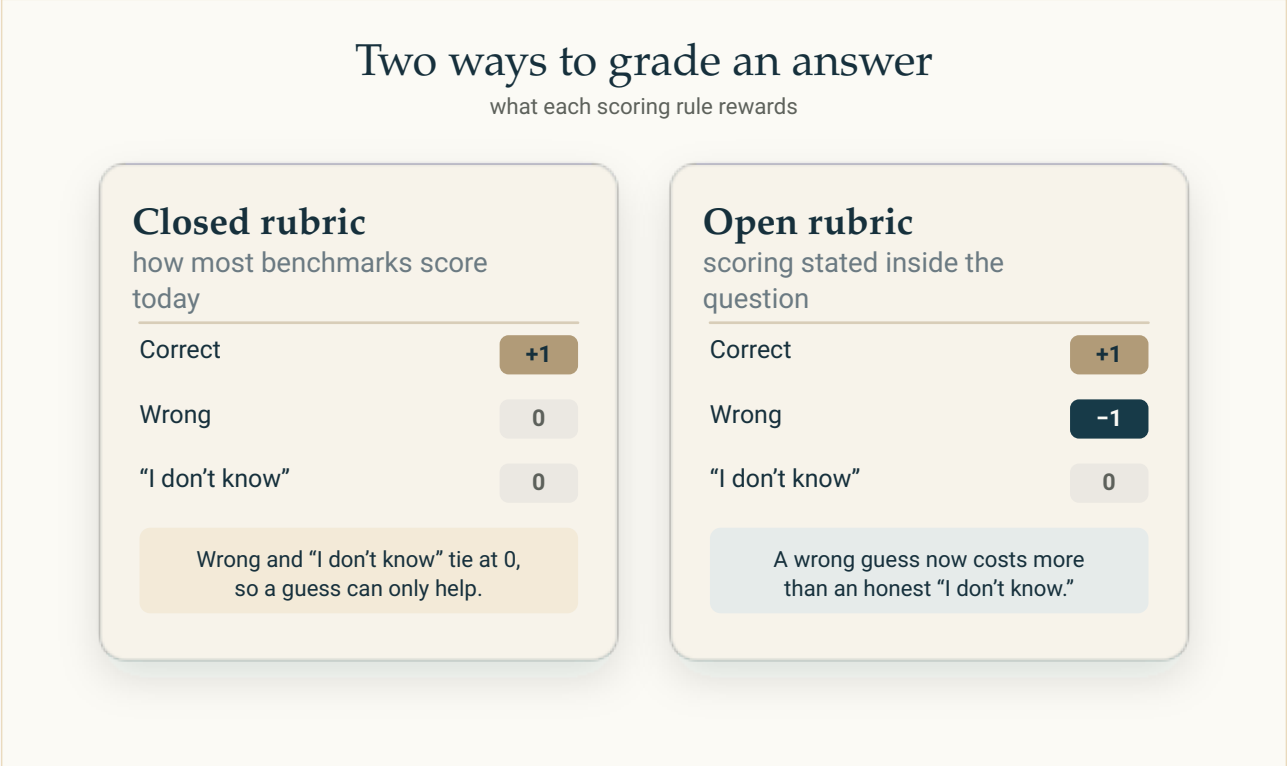
Turing'in aslında, bambaşka bir problemde sorduğu soru şuydu: *bir sonraki topun daha önce hiç görmediğiniz bir renkte olma olasılığı nedir?* Hiç çekmediğiniz şeyleri doğrudan sayamazsınız — ama yalnızca bir kez gördüğünüz renkleri, yani "singleton"ları, sayabilirsiniz. **Good–Turing tahmini** denen yöntem, örneğinizde yalnızca bir kez görülenlerin payının, henüz hiç görmediğiniz renklerde saklı toplam olasılığı tahmin ettiğini söyler. Yüz çekilişinizin onu birer kez görülen renklerdi; demek ki sıradaki topun yepyeni bir renk olma olasılığı yaklaşık **10 / 100 = %10**.

Bu bir kez görülen renkler *hata* değildir. Kendi bilgisizliğinizin ölçüsüdür: çok sayıda rengin yalnızca bir kez çıkması, örneklemin size dünyada henüz çekmediğiniz daha pek çok şey olduğunu söyleme biçimidir.

Şimdi renklerin yerine doğum tarihlerini, torbanın yerine de modelin eğitim metnini koyun. Diyelim ki modelin gördüğü doğum tarihlerinin **beşte biri yalnızca bir kez** geçiyor. Aynı numara: olasılık kütlesinin yaklaşık beşte biri, modelin fiilen hiç görmediği tarihlerde

> saklıdır — ve bir doğum tarihinin arkasında geri dönüp hesaplayabileceğiniz bir örüntü yoktur (bir insanın doğum gününü *çıkaramazsınız*). Bu nedenle bir kez görülen ya da hiç görülmeyen bir tarih, modelin ağırlıklandıramadığı bir yazı-tura gibidir ve bunların yaklaşık **beşte birinde** yanlış yapacaktır. Hiçbir zekâ bunu düzeltemez; öğrenilecek veri yoktu.
> Argümanın küçültülmüş hâli budur: singleton oranı, dünyanın ne kadarının *bu veriden* öğrenilemeyeceğini ölçer ve bu da hataların altına bir **taban** koyar. Bir modelin başkentleri neredeyse hiç kaçırmamasının nedeni de budur — *Paris* sürekli karşınıza çıkar, singleton oranı sıfıra yakındır ve öğrenilecek çok şey vardır.

Bu, halüsinasyonların nereden geldiğini açıklar. Ama neden *hayatta kaldıklarını* — yani modellerin daha sonra onları yardımsever ve dürüst yapmak amacıyla yapılan tüm eğitimlere rağmen neden hâlâ blöf yapıp şüphelerini itiraf etmediklerini — açıklamaz. Makalenin buradaki benzetmesi rahatsız edici ölçüde yerinde. Bir sınavda cevabı bilmeyen bir öğrenciyi düşünün. Boş bırakmak sıfır puansa, tahmin etmek ise bir puan getirebilirse, notu maksimize eden davranış tahmin etmektir — hem de kararlı ve spesifik biçimde, asla "emin değilim" demeden. Öğrenciler bunu öğrenir. Görünüşe göre modeller de öğreniyor — çünkü onları aynı şekilde notlandırıyoruz. Yazarlar, alanın gerçekten rekabet ettiği benchmark'ları, modellerin çıkmak için ayarlandığı liderlik tablolarını incelemiş ve neredeyse hepsinin "bilmiyorum" cevabına yanlış cevapla tamamen aynı puanı verdiğini görmüş: sıfır. Bu kural altında her zaman tahmin yapan bir model, belirsizliğini dürüstçe işaretleyen tamamen aynı bir modeli yener. Oldukça gerçek bir anlamda, onları bu davranışa puanlayarak itiyoruz.



Benchmark'lar tahmin yürütmeyi rasyonel hâle getirebilir. Çoğu benchmark'ın kullandığı puanlamada (solda), yanlış cevap ile dürüst bir "bilmiyorum" aynı şekilde sıfır puan alır — bu nedenle tahmin etmek yalnızca yardımcı olabilir. Kuralları sorunun içinde açıkça yazın ve yanlış cevaba ceza verin (sağda); o zaman emin olmadığında çekimser kalmak daha iyi hamle olabilir. Bu, testin neyi ödüllendirdiğini değiştirir; tek başına halüsinasyon

sorununu çözmez.

Original diagram — The Clean Paper CC BY 4.0

Akılda tutulmaya değer kısım bu, çünkü alışılagelmiş başlıkların tersine gidiyor. Halüsinasyon sıklıkla teknolojinin kaçınılmaz, neredeyse mistik bir sınırı gibi sunuluyor. Makale iki iddiaya da itiraz ediyor. Ön-eğitimdeki hata tabanı gizemli değildir — makine öğrenmesinin onlarca yıldır bildiği sıradan istatistiksel hatadır. Bu hataların ısrarla sürmesi de kaçınılmaz değildir — kısmen bizim kurduğumuz ve değiştirebileceğimiz bir teşviktir. Emin olmadığında basitçe cevap vermeyen bir sistem hiç halüsinasyon yapmazdı; konuşlandırılmış modellerin böyle davranmamasının nedenlerinden biri, puan tablolarımızın bu reddi cezalandırmasıdır.

Yazarlar ne yaptı?

Makalenin üç bölümü var. İlk olarak, "yalnızca geçerli metin üretme" probleminin en az ikili bir "bu ifade geçerli mi?" sınıflandırma problemi kadar zor olduğunu göstererek ön-eğitim sırasında belirli miktarda halüsinasyonun istatistiksel olarak zorunlu olduğuna ilişkin matematiksel bir argüman sunuyor. İkinci olarak, etkili on benchmark'ın incelenmesiyle desteklenen bir sav geliştiriyor: yaygın doğruluk türü metrikler, çekimser kalmak yerine tahmin yürütmeyi ödüllendiriyor. Üçüncü olarak bir düzeltme öneriyor ve bunu bir vaka çalışmasıyla deniyor: **açık değerlendirme kuralları (open rubrics)**. Burada puanlama kuralı doğrudan sorunun içinde belirtiliyor (örneğin "doğru cevap 1, yanlış cevap −1 puan; %50'den az eminsen çekimser kal"), böylece model dürüstlüğün ne zaman ödüllendirildiğini görebiliyor. Bunu dört frontier modelde — Google'ın Gemini 3 Pro'su, OpenAI'ın GPT-5'i, xAI'ın Grok 4'ü ve Anthropic'in Claude Opus 4.5'i — SimpleQA'nın 4.326 olgusal sorusuyla deniyorlar. Vaka çalışmasının yalnızca örnekleyici olduğunu, "modeller arasında kontrollü bir değerlendirme olmadığını" açıkça belirtiyorlar (varsayılan ayarlar, özel ayarlama yok, maliyet normalizasyonu yok).

Ne buldular?

- **Ön-eğitim belirli miktarda hatayı zorunlu kılıyor.** Bir modelin kendinden emin yanlışlar üretme oranı, ondan türetilen en iyi "bu ifade geçerli mi?" sınıflandırıcısının hata oranının (kabaca iki katı) tarafından aşağıdan sınırlanıyor. Öğrenilebilir örüntüsü olmayan olgular için bu taban en az *singleton oranı* — eğitimde tam bir kez görünen olguların payı. Veri tamamen temiz olsa bile bir miktar halüsinasyon kaçınılmaz.
- **Puanlama tahmin etmeyi somut biçimde ödüllendiriyor.** Sıradan doğru/yanlış puanlamada asla çekimser kalmamak optimal stratejidir ve yazarların incelemesinde popüler benchmark'ların büyük çoğunluğu "bilmiyorum" cevabını doğrudan yanlış sayıyor. Kendi örneklerinden çarpıcı biri: SimpleQA testinde ham doğruluk, hemen her soruyu cevaplayan ve dört cevaptan üçünden fazlasında yanlış olan OpenAI o4-mini'yi, emin olmadığında çekimser kaldığı için çok daha az hata yapan GPT-5-mini'nin biraz *önüne* koyuyor. Daha pervasız model puan tablosunda daha iyi görünüyor.
- **Açık kurallar teşviki tersine çeviriyor (vaka çalışmalarında).** Basit bir halüsinasyon azaltma

yöntemini test ediyorlar: modele iki kez cevap verdirip cevaplar uyuşmazsa çekimser kalmak. Standart doğruluk altında yöntem hataları azaltıyor *ama doğruluğu da düşürüyor* — dolayısıyla metrik yöntemin benimsenmesini caydırıyor. Açık kurallar altında ise aynı yöntem, bir dizi ceza seviyesinde dört modelin tümünde daha iyi çıkıyor; ayrıca ham doğruluğun emin olmadığında çekimser kaldığı için cezalandırdığı GPT-5-mini, puanlama açıkça belirtildiğinde o4-mini'nin önüne geçiyor (model başına n = 4.326 soru).

Bu büyük olasılıkla ne anlama geliyor?

Halüsinasyonu azaltmak, büyük ölçüde halüsinasyona özel daha fazla test icat etmekten çok, ana akım benchmark'ların belirsizliği nasıl puanladığını değiştirmekle ilgili. "Bilmiyorum" demek artık cezalandırılmadığında, dürüst belirsizlik avantaj hâline gelebilir. Puan tablosu değişmedikçe halüsinasyonu azaltmak modellerin doğruluk puanına mal olmaya ve dolayısıyla caydırılmaya devam edecek — yazarların sorunu "sosyo-teknik" diye adlandırmasının nedeni bu: biraz daha iyi metrik, biraz da etkili liderlik tablolarının bu metriği benimsemesi.

Bu neyi *kanıtlamıyor*?

- Açık kuralların gerçek dünyada halüsinasyonu düzelttiğini **göstermiyor**. Destekleyici deney, bilerek **kontrolsüz** bırakılmış küçük bir vaka çalışması — varsayılan ayarlarda dört model, seçilmiş tek bir azaltma yöntemi, tek bir olgusal soru-cevap testi. Amaç modelleri sıralamak ya da genel etkinliği kanıtlamak değil, *teşvikin tersine döndüğünü* göstermek.
- Puanlamanın *tek* neden olduğunu **iddia etmiyor**. Eğitim verisindeki hatalar, gerçekten zor problemler ve modele yabancı istemler ayrı kaynaklar olmaya devam ediyor.
- Halüsinasyonların kaçınılmaz olduğu yönündeki popüler söylemi **desteklemiyor**. Yazarlar tersini savunuyor: yalnızca doğrulanabilir soruları cevaplayan ve diğerlerinde "bilmiyorum" diyen bir sistem hiç halüsinasyon yapmazdı.
- Ön-eğitimdeki hata tabanını **ortadan kaldırmıyor** — onu açıklıyor ve sınırlandırıyor; üstelik bu sınır tüm model davranışlarını değil, kendinden emin olgusal hataları ilgilendiriyor.
- Açık kuralların tek başına *yeterli* olduğunu **göstermiyor**. Bir değerlendirmenin neyi ödüllendirdiğini değiştiriyorlar; retrieval, araç kullanımı veya daha iyi kalibre edilmiş modellerin yerini tutmuyorlar.

Kanıt ne kadar güçlü?

- Çekirdek kısım **matematik** — ölçüm değil, biçimsel alt sınırlar. Teorik bir argüman olarak kendi varsayımları altında sağlam.
- Problemin **bilerek basitleştirilmiş modellerine** dayanıyor; yazarlar her cevabı doğru, yanlış veya "bilmiyorum" olarak üçe ayırmanın "yanlış üçleme" olduğunu ve en temiz sınır için kullanılan

"keyfî olgular" senaryosunun idealize edildiğini kendileri belirtiyor.
- Benchmark incelemesi **küçük ve seçilmiş bir örneklem** — etkili on değerlendirme; kapsamlı bir audit değil.
- Vaka çalışması **gerçek ama sınırlı**: dört frontier model, tek bir azaltma yöntemi, yalnızca SimpleQA, varsayılan ayarlar ve açıkça "kontrollü bir değerlendirme değil." Teşvik argümanının kavram kanıtı; benchmark sonucu değil.
- Bakış açısını da adlandırmak gerekir: dört yazardan üçü **OpenAI** çalışanı veya eski çalışanı ve makale alanın modelleri nasıl değerlendirdiğini değiştirmesi gerektiğini savunuyor. Bu, çıkarı olan bir taraftan gelen iyi kurulmuş bir argümandır; tarafsız dış değerlendirme değildir — göz ardı etmek için değil, tartarken bilmek için. (Makale, eleştirisini başkalarına olduğu kadar kendi modelleri o4-mini ve GPT-5-mini'ye de yöneltiyor.)

Neden önemli?

Aşırı abartılan bir problemi yeniden çerçeveliyor. "Halüsinasyon" genellikle ya tekinsiz bir kusur ya da aşılamaz bir duvar gibi satılıyor; bu makale onu sıradan ve kısmen kendi yarattığımız bir sonuç hâline getiriyor — anlayabildiğimiz istatistiksel bir taban, üzerine bizim seçtiğimiz bir teşvik eklenmiş. Daha geniş ders daha sessiz ama daha kullanışlı: güvenilirlikte ilerleme, ne inşa ettiğimiz kadar *neyi ölçtüğümüze* de bağlı olabilir.

Kısa özet

Dil modellerinin kendinden emin yanlış cevapları iki kaynaktan geliyor. Birincisi istatistiksel: bir olgunun öğrenilecek örüntüsü yoksa, dili taklit etmek üzere eğitilmiş model bazen onu yanlış yapacaktır ve bu hata tabanı tahmin edilebilir (örneğin eğitimde kaç olgunun yalnızca bir kez geçtiğine bakarak). İkincisi teşvikler: modellerin sıralandığı neredeyse her benchmark "bilmiyorum" cevabını yanlış cevapla aynı puanladığı için tahmin etmek her zaman kazançlıdır — hatta soruların dörtte üçünden fazlasında yanlış yapan bir model, çekimser kalan daha dürüst bir modeli geçebilir. Yazarların önerisi yeni bir halüsinasyon testi değil, "open rubrics": puanlama kuralını sorunun içinde belirtmek. Dört frontier model üzerinde yapılan vaka çalışmasında bu, teşviki tersine çeviriyor ve halüsinasyonu azaltan bir yöntemi cezalandırmak yerine ödüllendiriyor. Bu, küçük ve açıkça kontrolsüz bir deneyle desteklenen teori + inceleme makalesi ve *Nature*'da hakemli olarak yayımlandı; önerilen düzeltmenin ölçekli biçimde işe yaradığı henüz gösterilmiş değil, ama halüsinasyonların ne gizemli ne de katı biçimde kaçınılmaz olduğu savunuluyor.

Abartısız değerlendirme

Makalenin gösterdiği: Ön-eğitim sırasında bazı halüsinasyonların zorunlu olduğu bir matematiksel alt sınır (örüntüsüz olgular için en az "singleton oranı"); önde gelen benchmark'ların çoğunun

“bilmiyorum” cevabına puan vermediğini gösteren bir inceleme; ve puanlamanın istemde açıkça belirtilmesinin (“open rubrics”), sıradan doğruluğun cezalandırdığı bir halüsinasyon azaltma yöntemini dört modelde ödüllendirdiği bir vaka çalışması.

Makul ama kanıtlanmamış olan: Açık kuralların ana akım benchmark’lara yerleştirilmesinin, konuşlandırılmış modellerde halüsinasyonu anlamlı ölçüde azaltacağı. Destekleyici deney küçük ve açıkça kontrolsüz.

Göstermediği: Halüsinasyonların kaçınılmaz olduğu (makale tersini savunuyor); puanlamanın tek neden olduğu; halüsinasyonun bütünüyle ortadan kaldırılabileceği; vaka çalışmasının dört modeli birbirine göre sıraladığı.

Başlıca sınırlamalar: Bilerek basitleştirilmiş doğru/yanlış/“bilmiyorum” modeli (yazarlar buna “yanlış üçleme” diyor); küçük ve seçilmiş benchmark incelemesi (on değerlendirme); kontrolsüz vaka çalışması (dört model, tek yöntem, tek test, varsayılan ayarlar); ayrıca alanın modelleri nasıl değerlendirmesi gerektiğine dair OpenAI ağırlıklı bir argüman.

Genel bir okur ne kadar güven duymalı? Halüsinasyonun ne gizemli ne de katı biçimde kaçınılmaz olduğuna ve ana akım benchmark’ların bugün tahmin yürütmeyi ödüllendirdiğine yüksek güven. Önerilen çözümün işe yaradığına orta düzeyde güven — artık gerçek bir kavram kanıtı var, fakat geniş ölçekte ve farklı koşullarda çalıştığını gösteren kanıt henüz yok.

Kaynaklar

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