The Clean Paper

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

Mengapa model bahasa berhalusinasi — dan mengapa cara kita memberi markah mengekalkannya begitu

Kepalsuan yakin yang kita panggil "halusinasi" bukan kecacatan misteri: sebahagiannya hasil sampingan statistik latihan, dan ia berterusan kerana penanda aras arus perdana memberi ganjaran kepada tekaan yakin berbanding "saya tidak tahu" yang jujur. Satu kajian kes pada empat model terdepan menunjukkan bahawa menyatakan aturan pemarkahan dalam prompt ("rubrik terbuka") membalikkan insentif itu.

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

Mengapa model meneka, dan mengapa kita mengajarnya berbuat demikian

Tanya sebuah model bahasa besar tarikh lahir orang yang tidak dikenali dan ia mungkin menjawab "7 Mac" dengan keyakinan tenang seperti sedang membacanya daripada kad — lalu salah, tiga kali berturut-turut, dengan tiga tarikh berbeza. Para penulis memberikan contoh tepat seperti ini: model terkemuka ditanya soalan fakta biasa — tarikh lahir seseorang, atau kepanjangan akronim yang jarang digunakan — setiap satu dengan yakin mereka jawapan berbeza, dan semuanya salah. Industri memanggilnya *halusinasi*, istilah yang membuatnya kedengaran seperti kecacatan persepsi. Langkah pertama kertas ini ialah menanggalkan misteri daripadanya.

Mulakan dengan cara model dibina. Dalam peringkat latihan pertama dan terbesarnya, ia pada dasarnya belajar rupa bahasa yang lancar dengan membaca jumlah teks yang sangat besar. Sekarang ambil satu fakta tanpa corak di belakangnya — tarikh lahir seorang individu tertentu. Jika tarikh itu muncul sekali dalam teks latihan, atau langsung tidak, tiada apa untuk dipaut oleh pembelajar corak: jawapannya, dari sudut pandang model, arbitrari. Para penulis menjadikan ini tepat dengan meminjam idea lama (daripada Alan Turing, untuk masalah lain): jika satu daripada lima tarikh lahir hanya muncul sekali dalam data, model sepatutnya dijangka salah pada sekurang-kurangnya satu daripada lima — bukan kerana ia rosak, tetapi kerana memang tiada apa di situ untuk dipelajari. (Dengan logik sama, model hampir tidak pernah tersilap ibu negara sesebuah negara: fakta itu muncul sentiasa.) Mereka berhujah dengan teliti bahawa membezakan kenyataan benar daripada kenyataan palsu yang munasabah itu sendiri masalah sukar, dan menghasilkan *hanya* kenyataan benar sekurang-kurangnya sama sukarnya. Ada lantai ralat yang memang terbina di dalamnya.

Idea di bawahnya: bagaimana mengira apa yang belum anda lihat

Ini bersandar pada idea yang benar-benar cerdik — lebih tua daripada model bahasa, dan patut dikenali dengan betul.

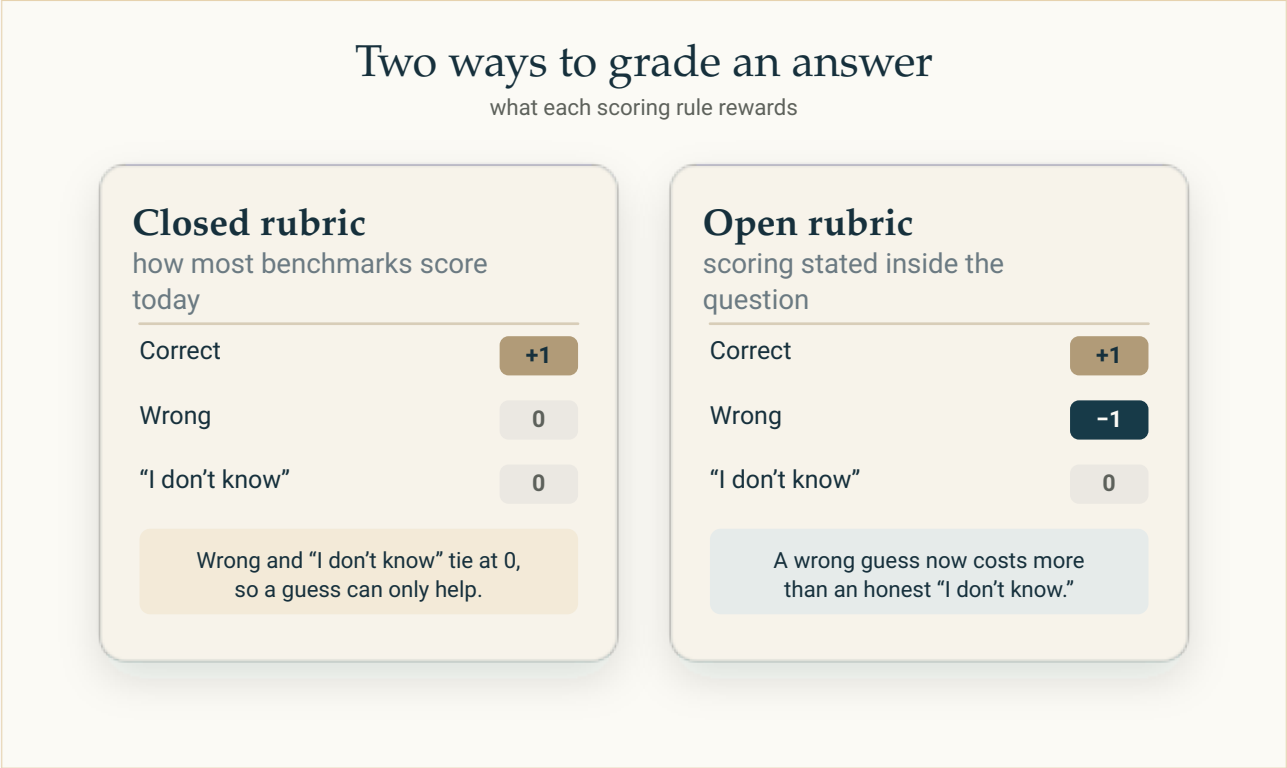
Mulakan dengan beg bola berwarna. Anda tidak tahu berapa banyak warna di dalamnya. Anda menarik 100 bola, satu demi satu, dan mencatatnya: merah 40, biru 25, hijau 15, kuning 5, ungu 3, oren 2 — kemudian **sepuluh warna berbeza yang masing-masing muncul tepat sekali.**

Sekarang soalan yang sebenarnya dihadapi Turing dalam masalah yang agak berbeza: *apakah kebarangkalian bola seterusnya mempunyai warna yang belum pernah anda lihat langsung?* Anda tidak boleh mengira perkara yang belum pernah ditarik — tetapi anda **boleh** mengira warna yang dilihat tepat sekali, iaitu "singleton". Helah yang dipanggil **anggaran Good–Turing** ialah bahagian cabutan yang merupakan singleton menganggarkan kebarangkalian yang masih tersembunyi dalam warna yang belum dilihat. Sepuluh daripada seratus cabutan anda ialah warna sekali sahaja, jadi peluang bola seterusnya ialah warna baharu kira-kira **10 / 100 = 10%.** Warna sekali lihat itu bukan *kesilapan*. Ia ukuran ketidaktahuan anda sendiri: banyak warna muncul sekali ialah cara sampel memberitahu bahawa dunia mengandungi lebih banyak warna yang anda belum sempat tarik.

Sekarang gantikan warna dengan tarikh lahir, dan beg dengan teks latihan model. Katakan, antara tarikh lahir yang dilihat, **satu daripada lima muncul tepat sekali.** Helah sama: kira-kira

> satu perlima kebarangkalian berada dalam tarikh lahir yang pada praktiknya tidak pernah dilihat model — dan tarikh lahir tiada corak untuk dijadikan sandaran (anda tidak boleh *mengira* tarikh lahir seseorang). Jadi tarikh yang dilihat sekali, atau tidak pernah, ialah syiling yang tidak dapat diberi wajaran oleh model, dan ia akan salah pada kira-kira **satu daripada lima**. Tiada kecerdikan boleh membaiki ini: memang tiada apa untuk dipelajari.
> Itulah keseluruhan hujah dalam bentuk kecil: kadar singleton mengukur berapa banyak dunia yang tidak boleh dipelajari *daripada data ini*, dan itu menjadi **lantai** di bawah ralat. Ia juga sebab model hampir tidak pernah tersilap ibu negara — *Paris* muncul berkali-kali, kadar singletonnya hampir sifar, jadi banyak yang boleh dipelajari.

Itu menerangkan dari mana halusinasi datang. Ia tidak menerangkan mengapa ia *kekal* — mengapa model, selepas semua latihan lanjut yang bertujuan menjadikannya membantu dan jujur, masih berpura-pura tahu daripada mengaku ragu. Di sini analogi kertas hampir terlalu tepat. Bayangkan pelajar dalam peperiksaan yang tidak tahu jawapan. Jika ruang kosong mendapat sifar dan tekaan mungkin mendapat satu, langkah yang memaksimumkan markah ialah meneka — dengan yakin, khusus, tidak pernah berkata "saya tidak pasti." Pelajar belajar begitu. Rupanya model juga — kerana kita memberi markah dengan cara sama. Para penulis meneliti penanda aras yang benar-benar dipertandingkan bidang ini, papan pendahulu yang cuba didaki model, dan mendapati hampir semuanya memberi "saya tidak tahu" tepat skor yang sama seperti jawapan salah: sifar. Di bawah aturan itu, model yang sentiasa meneka akan mengalahkan model serupa yang secara jujur menandakan ketidakpastian. Dalam erti yang agak literal, kita sedang memberi markah sehingga mereka terdorong berbuat demikian.



Penanda aras boleh menjadikan meneka rasional. Di bawah pemarkahan yang digunakan kebanyakan penanda

aras (kiri), jawapan salah dan “saya tidak tahu” yang jujur kedua-duanya mendapat sifar — jadi tekaan hanya boleh membantu. Nyatakan kaedah pemarkahan dalam soalan, dengan penalti bagi jawapan salah (kanan), dan memilih untuk tidak menjawab apabila ragu boleh menjadi pilihan lebih baik. Ia mengubah apa yang dihargai ujian; ia tidak, dengan sendirinya, menyelesaikan halusinasi.

Original diagram — The Clean Paper CC BY 4.0

Inilah bahagian yang patut dikekalkan kerana ia bertentangan dengan tajuk biasa. Halusinasi sering dijual sebagai had teknologi yang tidak dapat dielakkan, hampir mistik. Kertas ini menolak kedua-dua unsur itu. Lantai pralatihan bukan misteri — ia ralat statistik biasa, jenis yang difahami pembelajaran mesin selama beberapa dekad. Dan kelangsungannya bukan tidak dapat dielakkan — sebahagiannya ialah insentif yang kita bina dan boleh ubah. Sistem yang sekadar enggan menjawab apabila tidak pasti tidak akan berhalusinasi sama sekali; sebab model yang digunakan hari ini tidak berkelakuan begitu ialah papan markah menghukum penolakan.

Apa yang dilakukan oleh para penulis

Kertas mempunyai tiga bahagian. Pertama, hujah matematik bahawa sebahagian halusinasi dipaksa secara statistik semasa pralatihan, dengan menunjukkan bahawa “hasilkan hanya teks sah” sekurang-kurangnya sama sukar dengan masalah pengelasan binari “adakah kenyataan ini sah?” Kedua, hujah — disokong tinjauan sepuluh penanda aras berpengaruh — bahawa metrik gaya ketepatan arus perdana memberi ganjaran kepada meneka berbanding tidak menjawab. Ketiga, cadangan pembaikan dan kajian kes yang mengujinya: penilaian **rubrik terbuka**, di mana pemarkahan dinyatakan dalam soalan itu sendiri (contohnya, “jawapan betul mendapat 1, jawapan salah −1, jadi jangan jawab jika anda kurang daripada 50% pasti”), supaya model boleh mengetahui bila kejujuran diberi ganjaran. Mereka mencubanya pada empat model terdepan — Gemini 3 Pro Google, GPT-5 OpenAI, Grok 4 xAI dan Claude Opus 4.5 Anthropic — menggunakan 4,326 soalan fakta SimpleQA. Mereka secara jelas mengatakan kajian kes itu bersifat ilustratif, “bukan penilaian terkawal merentasi model” (tetapan lalai, tiada penalaan, tiada normalisasi kos).

Apa yang mereka temui

- **Pralatihan memaksa sebahagian ralat.** Kadar model mengeluarkan kepalsuan dengan yakin mempunyai had bawah kira-kira dua kali kadar ralat pengelas terbaik “adakah kenyataan ini sah?” yang dibina daripadanya. Bagi fakta tanpa corak boleh dipelajari, lantai itu sekurang-kurangnya *kadar singleton* — pecahan fakta yang muncul tepat sekali dalam latihan. Sebahagian halusinasi tidak dapat dielakkan walaupun data benar-benar bersih.
- **Pemarkahan memberi ganjaran kepada meneka — secara konkrit.** Di bawah pemarkahan betul/salah biasa, tidak pernah abstain ialah strategi optimum, dan tinjauan penulis mendapati sebahagian besar penanda aras popular menganggap “saya tidak tahu” hanya sebagai salah. Contoh jelas daripada model mereka sendiri: pada ujian SimpleQA, ketepatan mentah sedikit *memihak* kepada o4-mini OpenAI — yang menjawab hampir semuanya dan salah lebih tiga perempat masa

— berbanding GPT-5-mini, yang membuat jauh lebih sedikit kesilapan kerana abstain apabila tidak pasti. Model yang lebih berani kelihatan lebih baik di papan markah.
- **Rubrik terbuka membalikkan insentif (dalam kajian kes mereka).** Mereka menguji mitigasi halusinasi ringkas (minta model menjawab dua kali dan abstain jika dua jawapan tidak sepadan). Di bawah ketepatan piawai, mitigasi mengurangkan ralat *tetapi juga mengurangkan ketepatan* — jadi metrik menghalang penggunaannya. Di bawah rubrik terbuka, mitigasi sama menang bagi keempat-empat model merentasi julat penalti; dan GPT-5-mini — yang dihukum ketepatan mentah kerana abstain apabila ragu — mendahului o4-mini apabila kaedah pemarkahan dinyatakan secara terbuka (n = 4,326 soalan bagi setiap model).

Apa yang mungkin dimaksudkannya

Mengurangkan halusinasi kebanyakannya bukan perkara mencipta lebih banyak ujian khusus halusinasi. Ia tentang mengubah cara penanda aras arus perdana memberi markah kepada ketidakpastian supaya mengaku "saya tidak tahu" tidak lagi dihukum. Selagi papan markah tidak berubah, mengurangkan halusinasi akan terus mengorbankan mata ketepatan model dan oleh itu terus tidak digalakkan — sebab para penulis membingkai masalah sebagai "sosio-teknikal": sebahagian metrik yang lebih baik, sebahagian lagi meyakinkan papan pendahulu berpengaruh untuk menggunakannya.

Apa yang *tidak* dibuktikan oleh kajian ini

- Ia **tidak** menunjukkan rubrik terbuka menyelesaikan halusinasi di dunia sebenar. Eksperimen sokongan ialah kajian kes kecil yang sengaja **tidak terkawal** — empat model dengan tetapan lalai, satu mitigasi dipilih, satu ujian soal jawab fakta — bertujuan menunjukkan *pembalikan insentif*, bukan membuat kedudukan model atau membuktikan keberkesanan umum.
- Ia **tidak** mendakwa pemarkahan ialah *satu-satunya* punca. Ralat dalam data latihan, masalah yang benar-benar sukar dan prompt asing kekal sebagai sumber berasingan.
- Ia **tidak** menyokong slogan popular bahawa halusinasi tidak dapat dielakkan. Para penulis berhujah sebaliknya: sistem yang hanya menjawab soalan yang boleh disemak dan selainnya berkata "saya tidak tahu" tidak akan berhalusinasi.
- Ia **tidak** menghilangkan lantai pralatihan — ia menerangkan dan mengehadkannya, dan had itu menyangkut ralat fakta yakin, bukan semua tingkah laku model.
- Ia **tidak** menunjukkan rubrik terbuka *mencukupi* sendiri. Ia mengubah apa yang dihargai penilaian; ia bukan pengganti retrieval, penggunaan alat atau model yang tahap keyakinannya ditentukur dengan lebih baik.

Sejauh mana kukuh buktinya?

- Terasnya ialah **matematik** — had bawah formal, bukan ukuran. Sebagai hujah teori ia kukuh menurut andaian sendiri.
- Ia bersandar pada **model masalah yang sengaja dipermudah**; para penulis sendiri menandakan "trikotomi palsu" yang menganggap setiap respons sebagai betul, salah atau "saya tidak tahu," serta tetapan ideal "fakta arbitrari" yang digunakan untuk had paling bersih.
- Tinjauan penanda aras ialah **sampel kecil yang dipilih** — sepuluh penilaian berpengaruh, bukan audit menyeluruh.
- Kajian kes **nyata tetapi terhad**: empat model terdepan, satu mitigasi, SimpleQA sahaja, tetapan lalai, secara jelas "bukan penilaian terkawal." Ia bukti konsep untuk hujah insentif, bukan hasil penanda aras.
- Patut juga dinyatakan kedudukan penulis: tiga daripada empat penulis sedang atau pernah bekerja di **OpenAI**, dan kertas berhujah bidang perlu mengubah cara menilai model. Itu pendirian yang dihujahkan dengan baik daripada pihak berkepentingan, bukan ulasan luar yang neutral — patut ditimbang, bukan diketepikan. (Kredit kepada kertas: ia mengarahkan kritikan kepada model mereka sendiri, o4-mini dan GPT-5-mini, sama mudah seperti kepada yang lain.)

Mengapa ia penting

Ia membingkai semula masalah yang sangat digembar-gemburkan. "Halusinasi" cenderung dijual sama ada sebagai kecacatan aneh atau dinding yang tidak dapat dialihkan; kertas ini menjadikannya biasa dan sebahagiannya buatan kita sendiri — lantai statistik yang boleh difahami, di atas insentif yang kita pilih. Pelajaran lebih luas lebih senyap dan lebih berguna: kemajuan lanjut dalam kebolehpercayaan mungkin bergantung sama banyak pada *apa yang kita ukur* seperti pada apa yang kita bina.

Ringkasan utama

Jawapan palsu yang yakin daripada model bahasa datang daripada dua tempat. Pertama ialah statistik: apabila fakta tiada corak untuk dipelajari, model yang dilatih meniru bahasa kadangkala akan salah, dan lantai itu boleh dianggarkan (contohnya daripada berapa banyak fakta muncul sekali sahaja dalam latihan). Kedua ialah insentif: hampir setiap penanda aras tempat model diberi kedudukan memberi "saya tidak tahu" skor sama seperti jawapan salah, jadi meneka sentiasa menang — sehingga model yang salah tiga perempat masa boleh mengatasi model lebih jujur yang abstain. Cadangan penulis bukan satu lagi ujian halusinasi tetapi "rubrik terbuka": nyatakan kaedah pemarkahan dalam soalan. Dalam kajian kes empat model terdepan, itu membalikkan insentif supaya kaedah mengurangkan halusinasi diberi ganjaran dan bukannya dihukum. Ini kertas teori-plus-tinjauan dengan eksperimen kecil yang secara jelas tidak terkawal, telah disemak rakan sebaya dalam *Nature*; pembaikan itu menjanjikan tetapi belum ditunjukkan berfungsi pada skala besar, dan halusinasi

dihujahkan bukan misteri mahupun sepenuhnya tidak dapat dielakkan.

Semakan tanpa gembar-gembur

Apa yang ditunjukkan oleh kertas ini: Had bawah matematik di mana sebahagian halusinasi dipaksa semasa pralatihan (sekurang-kurangnya "kadar singleton" bagi fakta tanpa pola); tinjauan mendapati kebanyakan penanda aras utama tidak memberi kredit kepada "saya tidak tahu"; dan kajian kes empat model di mana menyatakan pemarkahan dalam prompt ("rubrik terbuka") membuat kaedah pengurangan halusinasi menang, manakala ketepatan biasa menghukumnya.

Apa yang munasabah tetapi belum dibuktikan: Bahawa rubrik terbuka, jika dimasukkan dalam penanda aras arus perdana, akan mengurangkan halusinasi secara bermakna dalam model yang digunakan. Eksperimen sokongan kecil dan secara jelas tidak terkawal.

Apa yang tidak ditunjukkannya: Bahawa halusinasi tidak dapat dielakkan (kertas berhujah sebaliknya); bahawa pemarkahan ialah satu-satunya punca; bahawa halusinasi boleh dihapuskan sepenuhnya; bahawa kajian kes memberi kedudukan empat model terhadap satu sama lain.

Had utama: Model betul/salah/"saya tidak tahu" yang sengaja dipermudah (penulis memanggilnya "trikotomi palsu"); tinjauan penanda aras kecil yang dipilih (sepuluh penilaian); kajian kes tidak terkawal (empat model, satu mitigasi, satu ujian, tetapan lalai); dan hujah yang dipimpin OpenAI tentang cara bidang patut menilai model.

Sejauh mana pembaca umum patut yakin? Tinggi bahawa halusinasi bukan misteri dan bukan mutlak tidak dapat dielakkan, dan bahawa penanda aras arus perdana kini memberi ganjaran kepada meneka. Sederhana bahawa pembaikan yang dicadangkan membantu — kini ada bukti konsep sebenar, tetapi belum demonstrasi bahawa ia berfungsi secara luas dan pada skala besar.

Sumber

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