The Clean Paper

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

Mengapa model bahasa berhalusinasi — dan mengapa cara kita menilainya membuat hal itu terus terjadi

Kepalsuan yang diucapkan dengan percaya diri dan kita sebut 'halusinasi' bukan gangguan misterius: sebagian merupakan produk sampingan statistik dari pelatihan, dan terus bertahan karena benchmark arus utama memberi keuntungan pada tebakan yakin dibanding 'saya tidak tahu' yang jujur. Studi kasus empat model frontier menunjukkan bahwa menyatakan aturan skor di dalam prompt ('open rubric') dapat membalik insentif tersebut.

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 menebak, dan mengapa kita mengajarinya untuk begitu

Tanyakan tanggal lahir orang yang tidak terkenal kepada sebuah model bahasa besar, dan ia mungkin menjawab "7 Maret" dengan keyakinan tenang seolah-olah sedang membacanya dari kartu — lalu salah, tiga kali berturut-turut, dengan tiga tanggal berbeda. Para penulis memberi contoh persis seperti ini: model-model terkemuka diberi pertanyaan faktual sederhana — tanggal lahir seseorang, atau kepanjangan sebuah akronim yang jarang dikenal — dan masing-masing dengan percaya diri mengarang jawaban berbeda, tanpa satu pun yang benar. Industri menyebutnya *hallucination* atau halusinasi, istilah yang membuatnya terdengar seperti gangguan persepsi. Langkah pertama makalah ini adalah menghilangkan aura misterius dari fenomena tersebut.

Mulailah dari cara sebuah model dibangun. Pada tahap pelatihan pertamanya dan yang terbesar, pada dasarnya model belajar seperti apa bahasa yang fasih dengan membaca teks dalam jumlah sangat besar. Sekarang ambil sebuah fakta yang tidak memiliki pola di baliknya — misalnya tanggal lahir satu orang tertentu. Jika tanggal itu muncul sekali dalam teks pelatihan, atau sama sekali tidak pernah muncul, tidak ada pola yang bisa dimanfaatkan oleh pembelajar pola: dari sudut pandang model, jawabannya arbitrer. Para penulis membuat gagasan ini presisi dengan meminjam ide lama (milik Alan Turing, untuk masalah yang berbeda): jika satu dari lima tanggal lahir hanya muncul sekali di data, maka model semestinya diperkirakan salah pada setidaknya satu dari lima tanggal tersebut — bukan karena model rusak, melainkan karena sejak awal tidak ada cukup informasi untuk dipelajari. (Dengan logika yang sama, model hampir tidak pernah salah menyebut ibu kota sebuah negara: fakta seperti itu muncul terus-menerus.) Mereka berargumen dengan hati-hati bahwa membedakan pernyataan benar dari pernyataan salah yang masuk akal sendiri merupakan masalah sulit, dan menghasilkan *hanya* pernyataan benar setidaknya sama sulitnya. Ada lantai kesalahan yang tertanam dalam persoalan.

Gagasan di baliknya: bagaimana menghitung apa yang belum pernah Anda lihat

Ini bertumpu pada ide yang benar-benar cerdik — lebih tua daripada model bahasa, dan layak dipahami dengan baik.

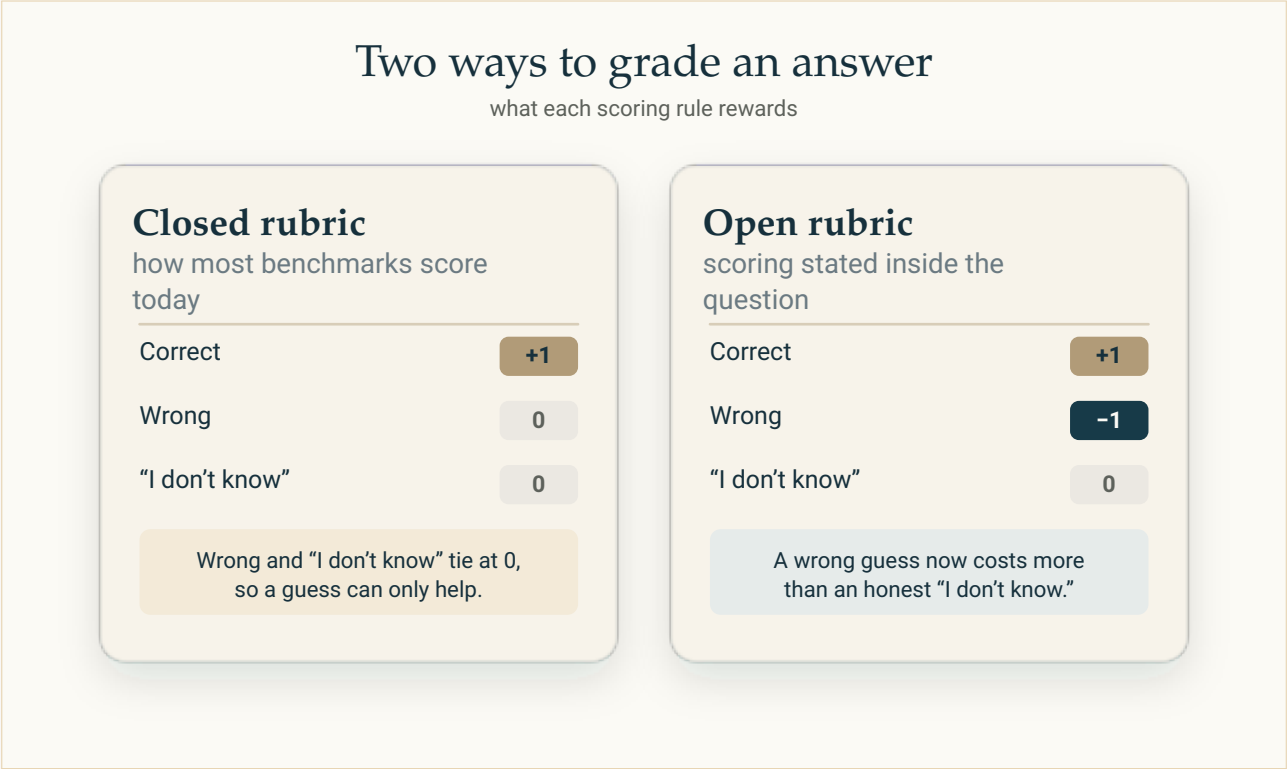
Mulailah dengan sekantong bola berwarna. Anda tidak tahu berapa banyak warna yang ada di dalamnya. Anda mengambil 100 bola satu per satu dan menghitung: merah 40, biru 25, hijau 15, kuning 5, ungu 3, jingga 2 — lalu **sepuluh warna berbeda yang masing-masing hanya muncul tepat sekali.**

Sekarang pertanyaan yang sebenarnya dihadapi Turing, dalam masalah yang sangat berbeda: *berapa peluang bola berikutnya memiliki warna yang sama sekali belum pernah Anda lihat?* Anda tidak dapat menghitung warna yang belum pernah terambil — tetapi Anda **dapat** menghitung warna yang hanya pernah muncul satu kali, yaitu "singleton". Trik yang disebut **estimasi Good–Turing** adalah bahwa proporsi pengambilan yang berupa singleton dapat memperkirakan probabilitas yang masih tersembunyi pada warna-warna yang belum terlihat. Sepuluh dari seratus bola Anda berasal dari warna yang hanya muncul sekali, jadi peluang bola berikutnya memiliki warna baru kira-kira **10 / 100 = 10%.**

Warna yang hanya muncul sekali itu bukan *kesalahan*. Mereka adalah ukuran ketidaktahuan

> Anda sendiri: banyak warna yang muncul sekali merupakan cara sampel memberi tahu bahwa dunia memiliki lebih banyak variasi yang belum sempat Anda ambil.
> **Sekarang ganti warna dengan tanggal lahir,** dan kantong dengan teks pelatihan model. Misalkan di antara tanggal lahir yang pernah dilihat model, **satu dari lima hanya muncul tepat sekali.** Triknya sama: kira-kira seperlima probabilitas berada pada tanggal lahir yang secara efektif belum pernah benar-benar dipelajari model — dan tanggal lahir tidak punya pola cadangan (Anda tidak dapat *menghitung* tanggal lahir seseorang dari prinsip lain). Jadi tanggal yang terlihat sekali, atau tidak pernah terlihat, adalah koin yang tidak dapat diberi bobot oleh model, dan model akan salah pada kira-kira **satu dari lima** kasus. Tidak ada kecerdikan yang dapat memperbaiki hal ini: memang tidak ada informasi yang bisa dipelajari.
> Itulah keseluruhan argumen dalam bentuk mini: tingkat singleton mengukur berapa banyak bagian dunia yang tidak dapat dipelajari *dari data ini*, dan itu menjadi **batas bawah** kesalahan. Itulah pula mengapa model hampir tidak pernah salah menyebut ibu kota — *Paris* muncul terus-menerus, tingkat singleton-nya mendekati nol, sehingga tersedia banyak informasi untuk dipelajari.

Itu menjelaskan asal halusinasi. Namun belum menjelaskan mengapa halusinasi *bertahan* — mengapa setelah semua tahap pelatihan lanjutan yang dimaksudkan agar model membantu dan jujur, model masih menggertak alih-alih mengakui keraguan. Di sini analogi makalah terasa hampir terlalu tepat. Bayangkan seorang siswa di ujian yang tidak tahu jawabannya. Jika jawaban kosong mendapat nilai nol dan tebakan mungkin mendapat satu poin, strategi yang memaksimalkan nilai adalah menebak — dengan percaya diri, spesifik, dan tidak pernah berkata “saya tidak yakin.” Siswa belajar melakukan ini. Ternyata model juga — karena kita menilainya dengan cara yang sama. Para penulis memeriksa tolok ukur yang benar-benar diperebutkan di bidang ini, papan peringkat yang menjadi sasaran penyetelan model, dan menemukan bahwa hampir semuanya memberi “saya tidak tahu” nilai yang sama persis dengan jawaban salah: nol. Dengan aturan itu, model yang selalu menebak akan mengalahkan model yang identik kecuali ia jujur menandai ketidakpastiannya. Dalam arti yang cukup harfiah, sistem penilaian kita mendorong mereka ke arah itu.



Sistem tolok ukur dapat membuat perilaku menebak menjadi strategi yang rasional. Dengan sistem penilaian yang dipakai sebagian besar tolok ukur (kiri), jawaban salah dan "saya tidak tahu" yang jujur sama-sama mendapat nol — jadi menebak hanya dapat membantu. Jika aturan dinyatakan di dalam pertanyaan, dengan penalti untuk jawaban salah (kanan), memilih tidak menjawab ketika ragu dapat menjadi strategi yang lebih baik. Ini mengubah apa yang dihargai oleh tes; dengan sendirinya, hal itu belum menyelesaikan halusinasi.

Original diagram — The Clean Paper CC BY 4.0

Bagian inilah yang paling layak dipertahankan, karena berlawanan dengan judul berita yang lazim. Halusinasi sering dijual sebagai batas teknologi yang tak terelakkan dan nyaris mistis. Makalah ini membantah kedua anggapan itu. Lantai kesalahan pada prapelatihan bukan misteri — itu kesalahan statistik biasa, jenis yang telah dipahami pembelajaran mesin selama puluhan tahun. Dan keberlanjutannya bukan sesuatu yang niscaya — setidaknya sebagian merupakan insentif yang kita bangun sendiri dan bisa kita ubah. Sistem yang cukup memilih tidak menjawab ketika tidak yakin tidak akan berhalusinasi sama sekali; alasan model yang digunakan sehari-hari tidak bertindak seperti itu adalah karena papan skor kita menghukum penolakan.

Apa yang dilakukan para penulis

Makalah ini terdiri dari tiga bagian. Pertama, argumen matematika bahwa sebagian halusinasi secara statistik dipaksa muncul selama prapelatihan, dengan menunjukkan bahwa "menghasilkan hanya teks yang valid" setidaknya sesulit masalah klasifikasi biner "apakah pernyataan ini valid?". Kedua, argumen — didukung survei terhadap sepuluh tolok ukur berpengaruh — bahwa metrik akurasi arus utama memberi insentif untuk menebak alih-alih abstain. Ketiga, usulan perbaikan dan studi kasus yang mengujinya: evaluasi **open-rubric**, yakni aturan penilaian dinyatakan di dalam pertanyaan itu

sendiri (misalnya, “jawaban benar mendapat 1, jawaban salah −1, jadi abstain jika keyakinan Anda kurang dari 50%”), sehingga model dapat mengetahui kapan kejujuran memang diberi nilai. Mereka mencobanya pada empat model frontier — Gemini 3 Pro dari Google, GPT-5 dari OpenAI, Grok 4 dari xAI, dan Claude Opus 4.5 dari Anthropic — menggunakan 4.326 pertanyaan faktual SimpleQA. Para penulis dengan jelas menyebut studi kasus ini ilustratif, “bukan evaluasi terkontrol lintas model” (pengaturan bawaan, tanpa tuning, tanpa normalisasi biaya).

Apa yang mereka temukan

- **Pretraining memaksa adanya sebagian kesalahan.** Tingkat model menghasilkan pernyataan salah dengan keyakinan tinggi memiliki batas bawah yang terkait dengan (kira-kira dua kali) tingkat kesalahan classifier terbaik “apakah pernyataan ini valid?” yang dibangun darinya. Untuk fakta tanpa pola yang dapat dipelajari, batas bawah itu setidaknya sebesar *singleton rate* — proporsi fakta yang hanya muncul satu kali selama pelatihan. Sebagian halusinasi tidak dapat dihindari bahkan jika datanya benar-benar bersih.
- **Sistem penilaian benar-benar memberi hadiah pada tebakan.** Dengan skor benar/salah biasa, tidak pernah abstain adalah strategi optimal, dan survei para penulis menemukan mayoritas besar tolok ukur populer memperlakukan “saya tidak tahu” sekadar sebagai jawaban salah. Contoh mencolok dari model mereka sendiri: pada SimpleQA, akurasi mentah sedikit *menguntungkan* o4-mini dari OpenAI — yang menjawab hampir semuanya dan salah lebih dari tiga perempat waktu — dibanding GPT-5-mini, yang membuat jauh lebih sedikit kesalahan karena memilih abstain ketika tidak yakin. Model yang lebih ceroboh terlihat lebih baik di papan skor.
- **Open rubric membalik insentif (dalam studi kasus mereka).** Mereka menguji mitigasi halusinasi sederhana: minta model menjawab dua kali dan abstain jika dua jawabannya berbeda. Dengan akurasi standar, mitigasi itu mengurangi kesalahan *tetapi juga mengurangi akurasi* — sehingga metrik justru tidak mendorong penerapannya. Dengan open rubric, mitigasi yang sama unggul untuk keempat model dalam berbagai tingkat penalti; dan GPT-5-mini — yang dihukum akurasi mentah karena abstain saat ragu — mengungguli o4-mini setelah aturan skornya dinyatakan secara terbuka (n = 4.326 pertanyaan per model).

Apa arti yang paling mungkin

Mengurangi halusinasi sebagian besar mungkin bukan soal menciptakan lebih banyak tes khusus halusinasi. Ini soal mengubah cara tolok ukur arus utama menilai ketidakpastian, sehingga mengakui “saya tidak tahu” tidak lagi dihukum. Selama papan skor tidak berubah, mengurangi halusinasi akan terus mengorbankan poin akurasi dan karena itu terus diberi insentif negatif — sebab itulah para penulis menyebut masalah ini “sosio-teknis”: sebagian soal metrik yang lebih baik, sebagian soal membuat papan peringkat berpengaruh benar-benar mengadopsinya.

Apa yang *tidak* dibuktikan

- Hasil ini **tidak** menunjukkan bahwa open rubric memperbaiki halusinasi di dunia nyata. Eksperimen pendukungnya adalah studi kasus kecil dan sengaja **tidak terkontrol** — empat model dengan pengaturan bawaan, satu mitigasi pilihan, satu tes tanya-jawab faktual — yang dimaksudkan untuk menunjukkan *pembalikan insentif*, bukan untuk memberi peringkat model atau membuktikan efektivitas umum.
- Hasil ini **tidak** mengklaim sistem penilaian adalah *satu-satunya* penyebab. Kesalahan dalam data pelatihan, masalah yang sungguh sulit, dan prompt yang asing tetap menjadi sumber terpisah.
- Hasil ini **tidak** mendukung kalimat populer bahwa halusinasi tidak dapat dihindari. Para penulis justru berargumen sebaliknya: sistem yang hanya menjawab pertanyaan yang dapat diverifikasi dan selain itu berkata "saya tidak tahu" tidak akan berhalusinasi.
- Hasil ini **tidak** menghapus lantai kesalahan prapelatihan — ia menjelaskan dan memberi batas padanya, dan batas tersebut berkaitan dengan kesalahan faktual yang diungkapkan dengan percaya diri, bukan seluruh perilaku model.
- Hasil ini **tidak** menunjukkan open rubric *cukup* dengan sendirinya. Ia mengubah apa yang dihargai sebuah evaluasi; ia bukan pengganti retrieval, penggunaan alat, atau model yang lebih terkalibrasi.

Seberapa kuat buktinya?

- Intinya adalah **matematika** — batas bawah formal, bukan pengukuran. Sebagai argumen teoretis, ia solid dalam asumsi yang digunakan.
- Argumen ini bertumpu pada **model masalah yang sengaja disederhanakan**; para penulis sendiri menandai "trikotomi palsu" yang memperlakukan setiap respons sebagai benar, salah, atau "saya tidak tahu", serta situasi ideal "fakta arbitrer" yang dipakai untuk batas paling bersih.
- Survei tolok ukur merupakan **sampel kecil yang dikurasi** — sepuluh evaluasi berpengaruh, bukan audit menyeluruh.
- Studi kasusnya **nyata tetapi terbatas**: empat model frontier, satu mitigasi, hanya SimpleQA, pengaturan bawaan, dan secara eksplisit "bukan evaluasi terkontrol". Ini bukti konsep untuk argumen insentif, bukan hasil tolok ukur komprehensif.
- Sudut pandangnya juga layak disebut: tiga dari empat penulis bekerja atau pernah bekerja di **OpenAI**, dan makalah ini berargumen bahwa bidang tersebut perlu mengubah cara mengevaluasi model. Itu posisi yang berargumentasi baik dari pihak yang berkepentingan, bukan tinjauan luar yang netral — perlu ditimbang, bukan diabaikan. (Untuk kreditnya, kritik tersebut diarahkan pada model OpenAI sendiri, o4-mini dan GPT-5-mini, sama mudahnya seperti ke model pihak lain.)

Mengapa ini penting

Makalah ini membingkai ulang masalah yang sangat dibesar-besarkan. “Halusinasi” cenderung dipasarkan sebagai cacat menyeramkan atau sebagai tembok yang tak dapat digeser; di sini ia dibuat lebih biasa dan sebagian merupakan akibat pilihan kita sendiri — lantai statistik yang benar-benar dapat dipahami, di atas insentif yang kita ciptakan. Pelajaran yang lebih luas lebih tenang dan lebih berguna: kemajuan lebih lanjut dalam keandalan mungkin bergantung sama besarnya pada *apa yang kita ukur* seperti pada apa yang kita bangun.

Singkatnya

Jawaban salah yang diberikan model bahasa dengan percaya diri berasal dari dua tempat. Yang pertama bersifat statistik: ketika sebuah fakta tidak memiliki pola yang dapat dipelajari, model yang dilatih meniru bahasa akan kadang salah, dan lantai kesalahan itu dapat diperkirakan (misalnya dari berapa banyak fakta yang hanya muncul sekali selama pelatihan). Yang kedua adalah insentif: hampir semua tolok ukur yang menentukan peringkat model memberi nilai “saya tidak tahu” sama dengan jawaban salah, sehingga menebak selalu menguntungkan — sampai-sampai model yang salah tiga perempat waktu dapat mengungguli model yang lebih jujur dan memilih abstain. Usulan para penulis bukan tes halusinasi baru, melainkan “open rubric”: nyatakan aturan penilaian di dalam pertanyaan. Dalam studi kasus empat model frontier, hal itu membalik insentif sehingga metode pengurang halusinasi diberi hadiah, bukan dihukum. Ini adalah makalah teori-plus-survei dengan eksperimen kecil yang secara eksplisit tidak terkontrol, telah melalui telaah sejawat di *Nature*; perbaikannya menjanjikan tetapi belum terbukti bekerja dalam skala luas, dan halusinasi dipandang bukan misterius maupun sepenuhnya tak terelakkan.

Penilaian tanpa berlebihan

Apa yang ditunjukkan makalah: Batas bawah matematis yang membuat sebagian halusinasi dipaksa muncul selama prapelatihan (setidaknya sebesar “singleton rate” untuk fakta tanpa pola); survei yang menemukan sebagian besar tolok ukur utama tidak memberi kredit pada “saya tidak tahu”; serta studi kasus empat model di mana menyatakan aturan skor di prompt (“open rubric”) membuat metode pengurang halusinasi unggul, padahal akurasi biasa justru menghukumnya.

Apa yang masuk akal tetapi belum terbukti: Bahwa open rubric, jika diadopsi dalam tolok ukur arus utama, akan secara berarti mengurangi halusinasi pada model yang digunakan sehari-hari. Eksperimen pendukung kecil dan secara eksplisit tidak terkontrol.

Apa yang tidak ditunjukkan: Bahwa halusinasi tak terelakkan (makalah ini berargumen sebaliknya); bahwa penilaian adalah satu-satunya penyebab; bahwa halusinasi dapat dihapus begitu saja; bahwa studi kasus ini memberi peringkat keempat model satu sama lain.

Keterbatasan utama: Model benar/salah/"saya tidak tahu" yang sengaja disederhanakan (para penulis menyebutnya "trikotomi palsu"); survei tolok ukur kecil dan terkurasi (sepuluh evaluasi); studi kasus tidak terkontrol (empat model, satu mitigasi, satu tes, pengaturan bawaan); serta argumen yang dipimpin OpenAI tentang bagaimana bidang tersebut seharusnya mengevaluasi model.

Seberapa besar keyakinan yang layak dimiliki pembaca umum? Keyakinan tinggi bahwa halusinasi bukan misteri dan juga tidak sepenuhnya tak terelakkan, serta bahwa tolok ukur arus utama saat ini memberi insentif untuk menebak. Keyakinan sedang bahwa perbaikan yang diusulkan akan membantu — kini ada bukti konsep nyata, tetapi belum ada demonstrasi bahwa ia bekerja 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