



The CLEAN PAPER

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

Me ya sa model na harshe suke yin hallucination —
da kuma me ya sa yadda muke ba su maki yake ci
gaba da haifar da shi

Karyar da ake fada da amincewa da muke kira “hallucination” ba wata matsala ta asiri ba ce: wasu sakamakon kididdiga na horo ne, kuma suna ci gaba domin manyan gwaje-gwaje suna ba zato mai karfin hali lada fiye da “ban sani ba” na gaskiya. Wani gwajin misali a kan manyan model huɗu ya nuna cewa faɗin ka’idojin maki a cikin umarni (“ka’idoji a fili”) yana juya wannan lada.

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

Me ya sa model suke yin zato, da kuma me ya sa muka koya musu haka

Ka tambayi babban model na harshe ranar haihuwar wani baƙo, zai iya amsa “Maris 7” da cikakkiyar amincewar wanda yake karantawa daga kati — amma ya yi kuskure sau uku a jere, da ranaku uku dabam-dabam. Marubutan sun bayar da irin wannan misali: an yi wa manyan model tambaya mai sauƙi ta gaskiya — ranar haihuwar wani ko cikakken ma’anar wata gajeriyaƙalma da ba a sani sosai ba — kowane ɗaya ya kirkiro wata amsa da amincewa, amma babu mai daidai. Kalmar da masana’antar ke amfani da ita ita ce *hallucination*, wadda take sa matsalar ta yi kama da kuskuren gani. Mataki na farko na makalar shi ne cire asirin daga ciki.

Mu fara da yadda ake gina model. A matakinsa na farko kuma mafi girma na horo, a aikace yana koyon yadda harshe mai kyau yake ta hanyar karanta rubutu mai yawa sosai. Yanzu mu ɗauki wata gaskiya da ba ta da tsarin da za a koya — ranar haihuwar wani mutum takamaimai. Idan ranar ta bayyana sau ɗaya kawai a rubutun horo, ko ba ta bayyana ba, babu wani tsari da mai koyon tsari zai rike: daga mahangar model, amsar ba ta da abin da zai nuna ta. Marubutan sun bayyana wannan daidai ta hanyar aron wata tsohuwar dabara (ta Alan Turing, daga wata matsala dabam): idan ranar haihuwa ɗaya cikin biyar ta bayyana sau ɗaya kawai a bayanar, ya kamata a sa ran samfurin ya yi kuskuren aƙalla ɗaya cikin biyar — ba domin ya lalace ba, sai domin babu abin da ya taɓa kasancewa a wurin da za a koya. (Da wannan tunani, model kusan ba sa kuskuren babban birnin ƙasa: irin waɗannan bayanai suna bayyana akai-akai.) Sun yi hujja da kulawa cewa bambance magana ta gaskiya daga ƙarya mai kama da gaskiya kansa wata matsala ce mai wuya, kuma samar da maganganun gaskiya *kawai* yana da wahala aƙalla kamar haka. Akwai wani matakin kuskure mafi ƙasa da ba za a kauce masa ba.

Tushen ra’ayin: yadda ake ƙidaya abin da ba ka gani ba tukuna

Wannan ya dogara da wata dabara mai wayo sosai — wadda ta girmi model na harshe kuma ta cancanci a fahimce ta da kyau.

Ka fara da buhun ƙwallaye masu launi. Ba ka san launuka nawa ne a ciki ba. Ka zaro 100 ɗaya bayan ɗaya ka ƙidaya: ja 40, shuɗi 25, kore 15, rawaya 5, shunayya 3, ruwan dorawa 2 — sannan **launuka daban-daban guda goma da kowannensu ya bayyana sau ɗaya kacal.**

Yanzu ga tambayar da Turing ya fuskanta a wata matsala dabam: *mene ne yiwuwar ƙwallo ta gaba ta kasance launin da ba ka taɓa gani ba?* Ba za ka iya ƙidaya abin da ba ka taɓa zaro ba — amma **za ka iya** ƙidaya launukan da ka gani sau ɗaya kawai, wato “singletons.” Dabarar da ake kira **Good-Turing estimation** tana amfani da rabon abin da ka zaro wanda singleton ne don kimanta yiwuwar da har yanzu take boye a launukan da ba ka gani ba. Goma cikin ɗari na abin da ka zaro launuka ne da suka bayyana sau ɗaya kawai, don haka yiwuwar ƙwallo ta gaba ta zama sabon launi gaba ɗaya kusan $10 / 100 = 10\%$.

Waɗannan launukan da suka bayyana sau ɗaya ba *kurakurai* ba ne. Ma’auni ne na abin da ba ka sani ba: idan launuka da yawa sun bayyana sau ɗaya, samfurin yana gaya maka cewa duniya tana da wasu da har yanzu ba ka zaro ba.

Yanzu ka musanya launuka da ranakun haihuwa, buhun kuma ya zama rubutun horon model. A ce daga ranakun da ya gani, **ɗaya cikin biyar ya bayyana sau ɗaya kawai.** Dabara ɗaya ce: kusan kashi ɗaya cikin biyar na yiwuwar yana cikin ranakun da a aikace model bai taɓa gani ba — kuma ranar haihuwa ba ta da tsarin da za a jingina da shi (ba za ka iya *lissafa* ranar haihuwar mutum ba). Don haka ranar da aka gani sau ɗaya ko ba a taɓa gani ba kamar tsabar da model ba zai iya auna ta ba ce, kuma zai yi kuskure a kusan **ɗaya cikin biyar.** Babu dabarar da za ta gyara wannan: babu abin da za a koya.

Wannan ita ce duk hujjar a takaice: yawan singleton yana auna nawa daga duniya ba za a iya koya *daga wadannan bayanai* ba, kuma hakan ya zama **matakin kasa** na kurakurai. Shi ya sa model kusan ba ya kuskuren babban birni — *Paris* yana bayyana akai-akai, yawansa na singleton ya kusa sifili, don haka akwai abin koyo sosai.

Wannan yana bayyana inda hallucination suke fitowa. Bai bayyana dalilin da ya sa suke *ci gaba* ba — me ya sa model, bayan duk horon da aka yi domin ya zama mai taimako da gaskiya, har yanzu yake yin karfin hali maimakon ya yarda da shakku. Misalin makalar a nan ya dace har ya sa mutum ya dan ji ba daɗi. Ka yi tunanin dalibi a jarabawa wanda bai san amsa ba. Idan barin fili ya samu sifili, zato kuma zai iya samu daya, abin da zai kara maki shi ne yin zato — da amincewa da takamaiman amsa, ba “ban tabbata ba.” Dalibai suna koyon haka. Model ma suna yi, domin haka muke ba su maki. Marubutan sun duba gwaje-gwajen da fannin yake fafatawa a kai da jerin matsayin da ake daidaita model domin su hau, suka ga kusan dukansu suna ba “ban sani ba” maki daya da amsar da ba daidai ba: sifili. A farkashin wannan ka’ida, model da ke yin zato koyaushe zai doke makamancinsa da ke faɗin rashin tabbas da gaskiya. A zahiri, hanyar ba da makinmu tana tura su yin haka.

Two ways to grade an answer

what each scoring rule rewards

Closed rubric

how most benchmarks score today

Correct	+1
Wrong	0
“I don’t know”	0

Wrong and “I don’t know” tie at 0, so a guess can only help.

Open rubric

scoring stated inside the question

Correct	+1
Wrong	-1
“I don’t know”	0

A wrong guess now costs more than an honest “I don’t know.”

Gwaje-gwaje na iya sa yin zato ya zama abin da ya dace. A tsarin makin da yawancin gwaje-gwaje suke amfani da shi (hagu), amsar da ba daidai ba da “ban sani ba” na gaskiya duka suna samun sifili — don haka zato zai iya taimakawa kawai. Idan aka faɗi ka’idoji a cikin tambayar, tare da hukunci ga kuskure (dama), kin amsawa idan ba a da tabbas zai iya zama zaɓi mafi kyau. Wannan yana sauya abin da gwajin yake ba lada; shi kaɗai ba ya warware hallucination.

Original diagram — The Clean Paper CC BY 4.0

Wannan shi ne ɓangaren da ya kamata a riƙe, domin ya saɓa wa kanun labari na yau da kullum. Sau da yawa ana sayar da hallucination a matsayin iyakar fasaha da ba za a kauce mata ba, mai kama da asiri. Makalar ta ƙi duk biyun. Matakin kuskuren horo na farko ba asiri ba ne — kuskuren kididdiga ne na yau da kullum, irin wanda koyon

na'ura ya fahimta shekaru da dama. Kuma ci gaba da kasancewarsa ba dole ba ne — a wani bangare, sakamakon abin da muka sa tsarin ya nema ne kuma za mu iya canjawa. Tsarin da kawai zai ki amsawa idan ba shi da tabbas ba zai yi hallucination ko ɗaya ba; dalilin da ya sa model da ake amfani da su ba sa haka shi ne jerin makinmu suna hukunta kin amsawa.

Abin da marubutan suka yi

Makalar tana da sassa uku. Na farko, hujjar lissafi cewa wasu hallucination suna zama dole ta fuskar kididdiga a horon farko, ta hanyar nuna cewa “samar da rubutu mai inganci kawai” yana da wahala aƙalla kamar matsalar rarrabawa ta “shin wannan magana ingantacciya ce?” Na biyu, hujja — tare da nazarin gwaje-gwaje goma masu tasiri — cewa ma'aunin da ke kama da accuracy na yau da kullum yana ba zato lada fiye da kin amsawa. Na uku, gyaran da aka gabatar da gwajin misali: kimantawa da **ka'idar maki a fili**, inda ake faɗin yadda za a ba maki a cikin tambayar kanta (misali, “amsa daidai tana samun 1, kuskure — 1, don haka ka ki amsawa idan tabbacin ka bai kai 50% ba”), domin samfurin ya san lokacin da ake ba gaskiya lada. Sun gwada wannan a kan manyan model huɗu — Gemini 3 Pro na Google, GPT-5 na OpenAI, Grok 4 na xAI da Claude Opus 4.5 na Anthropic — ta amfani da tambayoyin gaskiya 4,326 na SimpleQA. Sun bayyana a fili cewa gwajin misali ne, “ba kimantawa mai cikakken iko tsakanin model ba” (saitin asali, babu gyara, babu daidaita farashi).

Abin da suka gano

- **Horon farko yana tilasta wasu kurakurai.** Adadin da model yake fitar da karya da amincewa yana da wani iyaka na kasa da ya danganta da (kusan sau biyu) adadin kuskuren mafi kyawun mai rarraba “shin wannan magana ingantacciya ce?” da aka gina daga gare shi. Ga gaskiyar da ba ta da tsarin koyo, wannan iyakar aƙalla ita ce *yawan singleton* — rabon gaskiyar da ta bayyana sau ɗaya kawai a horo. Wasu hallucination ba za a kauce musu ba ko da bayanansu suna da tsabta sosai.
- **Ba da maki yana ba zato lada — a zahiri.** A tsarin daidai/kuskure na yau da kullum, rashin kin amsawa ko da yausha shi ne tsarin da ya fi kawo maki, kuma binciken marubutan ya gano yawancin shahararrun gwaje-gwaje suna ɗaukar “ban sani ba” a matsayin kuskure. Misali mai bayyana daga nasu gefen: a gwajin SimpleQA, accuracy na kai tsaye ya ɗan *fi son* o4-mini na OpenAI — wanda yake amsa kusan kome kuma yana kuskure fiye da kashi uku cikin huɗu na lokaci — a kan GPT-5-mini, wanda yake yin kurakurai kaɗan domin yana kin amsawa idan ba shi da tabbas. Model mai sakaci ya fi kyau a jerin maki.
- **Ka'idon da suke a fili suna juya abin da ake ba lada (a gwajin misalinsu).** Sun gwada wata hanya mai sauƙi ta rage hallucination (samsurin ya ba da amsa sau biyu, ya ki amsawa idan amsoshin biyu ba su yi daidai ba). A farkashin accuracy na yau da kullum, hanyar tana rage kurakurai *amma kuma tana rage accuracy* — don haka ma'aunin yana hana a rungume ta. A farkashin ka'idon da suke a fili, wannan hanyar ta fi kyau ga model huɗu duka a matakan hukunci daban-daban; GPT-5-mini kuma — wanda accuracy na kai tsaye ya hukunta saboda kin amsawa idan bai da tabbas — ya zarce o4-mini bayan an faɗi tsarin makin a fili ($n =$ tambayoyi 4,326 ga kowane model).

Abin da wannan mai yiwuwa yake nufi

Rage hallucination ba wai galibi batun kirkirar karin gwaje-gwaje na musamman gare shi ba ne. Batun sauya yadda manyan gwaje-gwaje suke ba rashin tabbas maki ne, domin yarda da “ban sani ba” kada ya ci gaba da zama laifi. Har sai jerin maki ya sauya, rage hallucination zai ci gaba da rage maki na accuracy, don haka za a ci gaba da hana shi — shi ya sa marubutan suka kira matsalar “ta zamantakewa da fasaha”: wani ɓangare shi ne inganta ma’auni, wani kuma samun manyan jerin matsayi su rungume shi.

Abin da wannan *bai tabbatar ba*

- **Bai nuna** cewa ka’idojin da suke a fili suna gyara hallucination a ainihin amfani ba. Gwajin da ke goyon bayansa karami ne kuma da gangan **ba a sarrafa shi cikakke ba** — model huɗu da saitin asali, hanyar rage matsala guda ɗaya, gwajin tambayar gaskiya guda ɗaya — an yi shi domin nuna *juya ladan*, ba domin tsara model ko tabbatar da tasiri gaba ɗaya ba.
- **Bai ce** ba da maki shi ne *kawai* dalili ba. Kurakuran bayanan horo, matsaloli masu wuya da gaske, da umarni da ba a saba da su ba suna zama wasu dalilai.
- **Bai goyi bayan** maganar da ake yadawa cewa hallucination ba za a kauce musu ba. Marubutan suna cewa akasin haka: tsarin da yake amsa tambayoyin da za a iya tabbatarwa kawai, ya ce “ban sani ba” ga sauran, ba zai taɓa yin hallucination ba.
- **Bai kawar** da matakin kuskure na horon farko ba — yana bayyanawa da iyakance shi, kuma iyakar tana magana ne kan kuskuren gaskiya da aka faɗa da amincewa, ba duk halayen model ba.
- **Bai nuna** cewa ka’idojin da suke a fili *sun isa* su kadai ba. Suna sauya abin da kimantawa yake ba lada; ba madadin nemo bayanai, amfani da kayan aiki, ko inganta daidaiton tabbacin model ba ne.

Yaya karfin shaidar yake?

- Tushen hujjar **lissafi** ne — iyakoki na kasa na ka’ida, ba ma’aunin gwaji ba. A matsayin hujjar ka’ida tana da karfi a karkashin sharuddanta.
- Ta dogara da **saukaka matsalar da gangan**; marubutan da kansu sun ambaci “rarrabuwar karya ta uku” da ke ɗaukar kowace amsa a matsayin daidai, kuskure, ko “ban sani ba,” da kuma yanayin “gaskiyar da ba ta da tsari” da aka tsara domin samun iyaka mafi tsafta.
- Nazarin gwaje-gwajen **karamin samfurin da aka zaɓa** ne — kimantawa goma masu tasiri, ba cikakken bincike ba.
- Gwajin misalin **na gaske ne amma mai iyaka**: manyan model huɗu, hanya ɗaya, SimpleQA kawai, saitin asali, kuma an bayyana shi “ba kimantawa mai cikakken iko ba.” Tabbacin ra’ayi ne kan lada, ba sakamakon gwajin kwatanta model ba.
- Ya dace a faɗi matsayin masu magana: marubuta uku cikin huɗu suna aiki ko sun taɓa aiki a **OpenAI**, kuma makalar tana neman fannin ya sauya yadda yake kimanta model. Hujja ce mai kyau daga ɓangaren da ke da ruwa a jallo, ba nazari mai zaman kansa daga waje ba — a auna ta, ba a yi watsi da ita ba. (Abin yabo shi ne makalar ta soki model nata, o4-mini da GPT-5-mini, kamar yadda ta soki sauran.)

Me ya sa yake da muhimmanci

Yana sauya yadda ake kallon matsalar da aka cika da hayaniya. Ana sayar da “hallucination” ko a matsayin wata matsala mai ban tsoro ko bango da ba ya motsi; makalar tana mayar da shi abu na yau da kullum kuma wani bangare muka jawo wa kanmu — matakin kasa na kididdiga da za mu iya fahimta, a saman abin da muka zaba mu ba lada. Babbar darasi ta fi shiru kuma ta fi amfani: karin ci gaba a amincin model na iya dogara da *abin da muke aunawa* kamar yadda yake dogara da abin da muke ginawa.

Takaitaccen bayani

Amsoshin karya da model na harshe suke bayarwa da amincewa suna fitowa daga wurare biyu. Na farko na kididdiga ne: idan gaskiya ba ta da tsarin koyo, model da aka horar ya kwaikwayi harshe zai yi kuskure wani lokaci, kuma za a iya kimanta wannan matakin kasa (misali daga adadin gaskiyar da ta bayyana sau daya kawai a horo). Na biyu shi ne lada: kusan kowane gwajin da ake tsara matsayin model da shi yana ba “ban sani ba” maki daya da amsar kuskure, don haka zato ya fi — har model da yake kuskure kashi uku cikin huɗu na lokaci zai iya zarce wanda ya fi gaskiya kuma yake ƙin amsawa. Shawarar marubutan ba wani sabon gwajin hallucination ba ne, sai “ƙa’idoji a fili”: a faɗi tsarin maki a cikin tambaya. A gwajin misali na manyan model huɗu, wannan yana juya ladan domin a ba hanyar rage hallucination maki maimakon a hukunta ta. Makala ce ta ƙa’ida da nazari tare da ƙaramin gwaji da aka bayyana ba a sarrafa shi ba, kuma an yi mata bitar masana a *Nature*; gyaran yana da alkawari amma ba a nuna yana aiki a faɗi da babban mataki ba, kuma hujjar ita ce hallucination ba asiri ba ne kuma ba lallai ba ne su zama abin da ba za a kauce masa ba.

Binciken ba tare da ƙarin gishiri ba

Abin da makalar ta nuna: Wani iyaka na kasa na lissafi wanda wasu hallucination suke zama dole a horon farko (akalla “yawan singleton” ga gaskiyar da ba ta da tsari); nazari da ya gano manyan gwaje-gwaje ba sa ba “ban sani ba” maki; da gwajin model huɗu wanda faɗin tsarin maki a umarni (“ƙa’idoji a fili”) ya sa hanyar rage hallucination ta yi nasara, inda accuracy na yau da kullum ya hukunta ta.

Abin da mai yiwuwa ne amma ba a tabbatar ba: Cewa shigar da ƙa’idoji a fili cikin manyan gwaje-gwaje zai rage hallucination sosai a model da ake amfani da su. Gwajin da ke goyon baya ƙarami ne kuma an bayyana ba a sarrafa shi ba.

Abin da ba ta nuna ba: Cewa hallucination ba za a kauce musu ba (makalar tana cewa akasin haka); cewa ba da maki shi ne dalili kaɗai; cewa za a iya kawar da hallucination gaba ɗaya; ko cewa gwajin misalin yana tsara model huɗu daga mafi kyau zuwa mafi muni.

Manyan iyakoki: Tsarin da aka saukaƙa da gangan na daidai/kuskure/“ban sani ba” (marubutan sun kira shi “rarrabuwar karya ta uku”); ƙaramin nazarin gwaje-gwajen da aka zaba (kimantawa goma); gwajin misali da ba a sarrafa shi ba (model huɗu, hanya ɗaya, gwaji ɗaya, saitin asali); da hujjar da masu OpenAI suka jagoranta kan yadda fannin ya kamata ya kimanta model.

Yaya karfin amincewar mai karatu na yau da kullum ya kamata ya kasance? Babbar amincewa cewa hallucination ba asiri ba ne kuma ba lallai ba ne su zama abin da ba za a kauce musu ba, kuma manyan gwaje-gwaje a yanzu suna ba zato lada. Matsakaicin amincewa cewa gyaran da aka gabatar yana taimakawa — yanzu yana da ainihin tabbacin ra'ayi, amma har yanzu ba a nuna yana aiki a fadɓi da babban mataki ba.

Majiyoyi

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>