



The CLEAN PAPER

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

Ìdí tí language model fi ń hallucinate — àti ìdí tí ọ̀nà tá a fi ń fún wọn ní máàkì fi ń mú un dúró

Àwọn èké tí a ń pè ní “hallucination” tí model ń sọ pẹ̀lú idánílójú kì í ẹ̀ àbùkù àdìtú: àwọn kan jẹ̀ èsì statistiki tí ikòni, wọn sì ń dúró nùtorí pé benchmark pàtàkì ń san èrè fún ìmẹ̀fò pẹ̀lú idánílójú ju “Mi ò mò” tó jẹ̀ òtútọ̀ lọ. Case study lórí frontier model méréń fi hàn pé síso ọ̀nà fífúnni ní máàkì nínú prompt (“open rubric”) yí ìwúrí náà padà.

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

Ìdí tí àwọn àwòṣe fi n méfò, àti ìdí tí a fi kọ wọn láti ẹ bẹẹ

Beèrè lówó large èdè àwòṣe nipa ojó ibí àjẹ̀jì kan, ó lè dàhùn “March 7” pẹ̀lú ìdánílójú bí ẹ̀ni tó n ká á lórí kààdì — kí ó sì ṣàṣiṣe lẹ́mẹ́ta ọ̀tọ̀ọ̀tọ̀ pẹ̀lú ojó mẹ́ta tó yàtò. Àwọn ònṁkòwé fúnni ní irú àpẹẹrẹ yí gan-an: wọn beèrè ibeèrè ọ̀títọ̀ lasán lówó àwọn àwòṣe aṣáájú — ojó ibí ènìyàn kan, tàbí ìtumọ̀ acronym kan tí a kò mọ̀ púpọ̀ — ọ̀kọọkan wọn sì dá èsì mǐ sílẹ̀ pẹ̀lú ìdánílójú, tí kò sí èyíkẹ́yí tó tọ̀. Ọ̀rọ̀ tí ilẹ̀-ìṣẹ̀ n ló fún èyí ni *hallucination*, tó mú un dà bí àṣiṣe nínú ohun tí a rí. Ìgbésẹ̀ àkókọ̀ iwé náà ni láti yọ àdìittú kúrò nínú rẹ̀.

Bèrè pèlú bí a ɛ kọ àwòṣe. Ní ìpele ìkòní àkòkọ́ tò sí tòbí jù lọ, ó kẹkẹ̀, ní kókó, bí èdè tò ní Ẹ̀n ɗáádáá ɛ rí nípá kíká ọ̀pọ̀lopọ̀ ọ̀rọ̀. Wá mú òtítọ́ kan tí kò ní àpẹrẹ̀ lẹ̀yìn rẹ̀ — ọ̀jọ̀ ìbí ènìyàn pàtó kan. Bí ọ̀jọ̀ nàà bá fara hàn nínú ọ̀rọ̀ ìkòní lẹ̀ẹ́kan, tàbí kò fara hàn rárá, kò sí ohun tí akẹ́kọ̀-àpẹrẹ̀ lè dì mú: láti ọ̀jọ̀ iwọ́ye àwòṣe, èsì nàà jẹ́ ohun tí kò ní ìlànà. Àwọn ònìkòwé sọ èyí di pípé nípá lílo èrò àtíjọ́ kan (tí Alan Turing, láti inú ìṣòro mún): bí ọ̀jọ̀ ìbí kan nínú mǎrùn-ún bá fara hàn lẹ̀ẹ́kan ɔ̀so nínú dátà, a yẹ kí a retí pé àwòṣe yóò ɔ̀sàṣe ó kéré tán lórí ọ̀kan nínú mǎrùn-ún — kì í ɛ nítorí pé ó bàjẹ́, sùgbọ́n nítorí pé kò sí ohun kan nìbẹ̀ láti kọ. (Pèlú ìrònú kan nàà, àwòṣe fẹ́rẹ́ má ɔ̀sàṣe lórí olú-ìlú orílẹ̀-èdè: wọn máa ní fara hàn ní gbogbo ìgbà.) Wọn jiyàn pèlú ìṣọ́ra pé yíya gbólòhùn òtítọ́ sọ̀tọ́ kúrò nínú èké tò ɛ́e gbà jẹ́ ìṣòro tò nira fúnra rẹ̀, àti pé pípèsè gbólòhùn òtítọ́ *nikan* ó kéré tán le tò ìṣòro yẹn. Ìpele àṣìṣe kan wà nínú ètò nàà láti ìbèrẹ̀.

Èrò tó wà lábẹ̀ rẹ̀: bí a ẹ̀e n ka ohun tí a kò tī rí

Èyí dúró lóri èrò ọlọgbọn gidi kan — tó dàgbà ju èdè àwòsè lọ, tó sì yẹ kí a mò dáadáa.

Bèrè pèlú àpò bọ̀lù aláwọ̀. O kò mọ́ iye àwọ̀ tó wà nínú rẹ̀. O fa 100 jáde, òkòòkan lẹ́sẹ́sẹ́, o sì ka wọ̀n: pupa 40, búlúù 25, ewé 15, yẹ̀lò 5, aláwọ̀ elése àlùkò 3, ọ̀sàn 2 — lẹ́yìn nàá **àwọ̀ mẹ̀wàá ọ̀tọ̀ọ̀tọ̀ tí ọ̀kọ̀ọ̀kan fara hàn lẹ̀ẹ̀kan soso.**

Wá wo ibèèrè tí Turing dojú kọ ní tòtótó, lórí isòro tò yàtò: *kí ni ànfààní pé bọ̀lù tò kàn jẹ àwò tí o kò tū rí rárá?* O kò lè ka ohun tí o kò tū fa jáde — sùgbón o **lè** ka àwọn àwò tí o tí rí lẹ́ẹ̀kan soṣo, àwọn “singleton.” Ogbón nàà, tí a n pè ní **Good–Turing estimation**, ni pé ìpín àwọn ifà rẹ̀ tò jẹ singleton nì fojú díwọn probability tò sì fara pamọ̀ sínú àwọn àwò tí o kò tū rí. Mewáá nínú ogòrùn-ún ifà rẹ̀ jẹ àwò tí o rí lẹ́ẹ̀kan soṣo, nítorí nàà ànfààní pé bọ̀lù tò kàn jẹ àwò tuntun pátápátá jẹ nńkan bí **10 / 100 = 10%**.

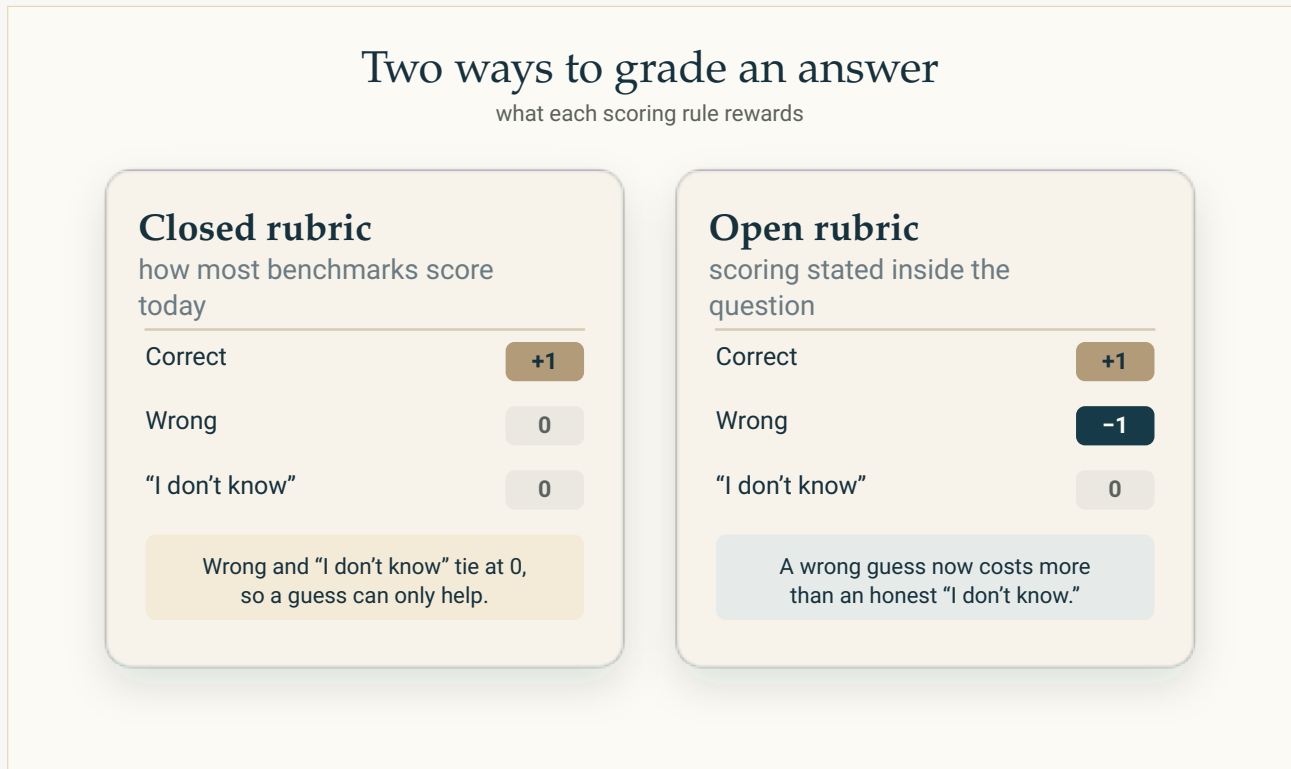
Àwọn àwò tí a rí lẹ́ẹ̀kan nàà kí í se àṣìṣe. Wọn jẹ ọ̀ṣùwọn àìmọ̀ tìrẹ̀: bí ọ̀pọ̀ àwò bá fara hàn lẹ́ẹ̀kan, àpẹẹrẹ nàà ní ọ fún ọ pé ayé ní púpọ̀ sí í tí ọ kò tī fa jáde.

Wá fi ojò ibi rópò àwò, kí o sì fi ọrò ìkọ̀nì àwòṣe rópò àpò nàà. Ká sọ pé, nínú àwọn ojò ibi tí ó rí, ọ̀kan nínú mǎrùn-ún fara hàn lẹ́ẹ̀kan soṣo. Ogbón kan nàà: ǹhkan bí ìdǎ mǎrùn-ún probability wà nínú àwọn ojò ibi tí àwòṣe kò rí ní tòótọ́ — ojò ibi kò sì ní àpẹẹrẹ́ tí a lè gbára lé (o kò lè *sírd* ojò ibi ẹ̀nì kan). Nítorí nàà, ojò tí a rí lẹ́ẹ̀kan, tàbí tí a kò rí rárá, dà bí owó tí àwòṣe kò lè fi ìwò̀n sí, yóò sì sàṣìṣe lórí ǹhkan bí ọ̀kan nínú mǎrùn-ún won. Kò sí ogbón tó lè tún èyí ṣe: kò sí ohun níbè láti kọ.

Èyí ní gbogbo àríyànjìyàn nàà ní kúkúró: iye singleton ní díwòṇ iye ayé tí a kò lè kọ *látí inú dátà yí*, èyí sì di **ààlà isàlẹ́** fún àṣìṣe. Ó tún jẹ́ ìdí tí àwòṣe fẹ́rẹ́ má ṣe ṣàṣìṣe olú-ilú kan — *Paris* fara hàn ní gbogbo ìgbà, iye singleton rẹ̀ sún mó òdò, nítorí nàà òpòlopò ohun ló wà látí kọ.

Eyí sàlàyé ibi tí hallucination ti wá. Kò sàlàyé idí tí wọn fi yè — idí tí àwọn àwòṣe, lẹyìn gbogbo ikòni tó tẹlẹ́ tí a se láti mú wọn wúlò kí wọn sì jẹ olóòtító, sì n se bí eni pé wọn mọ́ dípò kí wọn jẹwò àní-àní. Nihìn-ín, àfiwé iwé nàà fẹrẹ́ jẹ ohun tó n kó ènìyàn ní itìjù. Foju inú wo akẹkọ̀ọ́ nínú idánwò tí kò mọ èsì. Bí ààyè ọfọ bá gba ọdò, tí ìmẹfò sì lè gba ọ̀kan, ìgbèsè tó n mú máàkì pọ̀ jù ní láti mẹfò — pèlú idánílọ́jú àti kíkún, kì í se “Mi ọ dàjù.”

Àwọn akẹkọọ ní kọ èyí. Ó hàn pé àwọn àwòṣe náà ní kọ ọ — nítorí pé bá yí náà ní a ẹ ní fún wọn ní máàkì. Àwọn ònṁkòwé ẚà yèwò àwọn benchmark tí ẹka náà ní díje lórí gan-an, àwọn leaderboard tí a ní ẹtò àwòṣe láti gòkè, wọn sì rí i pé fẹrẹẹ gbogbo wọn ní fún “Mi ò mò” ní máàkì kan náà pèlú ẹ̀sì àṣìṣe: 0do. Lábẹ̀ òfin yẹn, àwòṣe tó máa ní mífò ní gbogbo ìgbà yóò borí àwòṣe tó jọ ọ ní gbogbo ọ̀nà mífù sùgbọ́n tó ní sọ àní-àní rẹ̀ ní òtítọ́. Ní ìtumọ̀ tó fẹrẹẹ jẹ̀ tààrà, a ní fí máàkì kọ wọn láti ẹ́.



Àwọn benchmark lè mú ìméfò jẹ́ ohun tó bójú mu. Lábẹ̀ ọ̀nà fífúnni ní máàkì tí ọ̀pọ̀ benchmark ń lò (òsì), ẹ̀sì àṣìṣe àti “Mi ò mò” tó jẹ́ òtítọ́ ń gba 0do — nítorí náà ìméfò lè ẚẹ̀rànwọ̀ nìkan. Sọ àwọn òfin náà nínú ìbèèrè, pèlú ìjìyà fún àṣìṣe (òtún), ìkọ̀ láti dáhùn nígbà àní-àní sì lè di ìgbésẹ̀ tó dára jù. Ó ń yí ohun tí ìdánwò ń san èrè fún padà; fúnra rẹ̀, kò yanjú hallucination.

Original diagram — The Clean Paper CC BY 4.0

Apá yí ló yẹ ká pa mó, nítorí pé ó lòdì sí àkólé tí a sáà rí. Wọn sáà máa ní ta hallucination gégé bí ààlà tí kò ẚeé yẹ, tó fẹrẹẹ jẹ́ ohun ìjìnlẹ̀ tí technology. Ìwé náà tako ìṣòrọ̀ méjèjì. Ààlà ìsàlẹ̀ pretraining kì í ẚe àdìtú — àṣìṣe statistiki lasán nì, irú èyí tí machine learning tí mò fún ọ̀pọ̀ ọ̀dún. Ìdúró rẹ̀ pé náà kò sì ẚe dandan — ní apá kan, ìwúrí tí àwa fúnra wa kọ ní, a sì lè yí i padà. Ètò kan tó kàn kọ̀ láti dáhùn nígbà tí kò dájú kò ní hallucinate rárá; ìdí tí àwọn àwòṣe tí a ní lò kò fí ẚe bẹ̀ẹ̀ ní pé scoreboard wa ní fí ìjìyà jẹ́ ìkọ̀ láti dáhùn.

Ohun tí àwọn ònṁkòwé ẚe

Ìwé náà ní apá mètá. Àkọ́kọ̀, àrìyànjiyàn mathematiki pé hallucination diẹ́ jẹ́ dandan ní statistiki nígbà pretraining, nípa fífi hàn pé “pèsè ọ̀rọ̀ tó fẹ́ṣe múlẹ̀ nìkan” ó kéré tán le tó ìṣòro classification alakomeji “ṣé gbólóhùn yí fẹ́ṣe múlẹ̀?” Èlẹ́kejì, àrìyànjiyàn — tí iwádìí benchmark olókíkí mewaà ẚe àtìlẹ́yìn fún — pé àwọn metric ipéye tó

wópò ní san èrè fún ìmèfò ju ìkò láti dáhùn lọ. Elèkẹta, àtúnṣe tí wọn dábàà àti ọ̀ràn iwádíí tó dán an wò: ìdánwò **open-rubric**, níbi tí a ti sọ ọ̀nà fífúnni ní máàkì nínú ìbèèrè fúnra rẹ (fún àpẹẹrẹ, “èsì tó tọ gba 1, èyí tó sàṣìṣe gba –1, nítorí nàà má dáhùn bí ìdánilójú rẹ kò bá tó 50%”), kí àwòṣe lè mọ ìgbà tí a ní san èrè fún òtítọ́. Wọn dán èyí wò lórí frontier àwòṣe méréin — Gemini 3 Pro ti Google, GPT-5 ti OpenAI, Grok 4 ti xAI àti Claude Opus 4.5 ti Anthropic — ní lílo àwọn ìbèèrè òtítọ́ 4,326 ti SimpleQA. Wọn sọ ní kedere pé ọ̀ràn iwádíí nàà jẹ àpẹẹrẹ, “kí í ṣe ìdánwò tó ní ìṣàkóso láàárín àwòṣe” (ètò default, kò sí tuning, kò sí ìṣòkan owó).

Ohun tí wọn rí

- **Pretraining ní fipá mú àṣìṣe díẹ.** Iye tí àwòṣe fi ní sọ èké pẹ̀lú ìdánilójú ní ààlà ìsàlẹ̀ tó jẹ́ (ní àfójúsùn, ilópo méjì) iye àṣìṣe classifier “ṣé gbólóhùn yí fẹ̀ṣẹ̀ múlẹ̀?” tó dára jù lọ tí a kọ́ láti inú rẹ̀. Fún àwọn òtítọ́ tí kò ní àpẹẹrẹ tí a lè kọ́, ààlà ìsàlẹ̀ nàà ó kéré tán jẹ́ *ìye singleton* — ipín àwọn òtítọ́ tí ó fara hàn lẹ́ẹ̀kan soṣo nínú ìkòni. Hallucination díẹ̀ kò ṣeé yẹ kòdà pẹ̀lú dátà tó mọ́ pátápátá.
- **Fífúnni ní máàkì ní san èrè fún ìmèfò — ní kedere.** Lábẹ́ fífúnni ní máàkì tó ní sọ pé ó tọ́ tàbí ó sàṣìṣe, àìkò láti dáhùn láé ni strategy tó dára jù, iwádíí àwọn ò̀nkòwé sì rí i pé ọ̀pọ́ jù lọ benchmark olókíkí ní ka “Mi ò mò” sí àṣìṣe lasán. Àpẹẹrẹ tó hàn gbangba láti ọ̀dọ́ wọn: lórí ìdánwò SimpleQA, raw ipéye díẹ̀ *se ojú rere sí* o4-mini ti OpenAI — tó dáhùn fẹ̀rẹ́ ohun gbogbo tí ó sì sàṣìṣe ju idámẹ́ta-méréin lọ — ju GPT-5-mini, tó ṣe àṣìṣe díẹ̀ púpọ́ nítorí pé ó kò láti dáhùn nígbà àní-àní. àwòṣe tó ní ṣe ewu jù ló dà bí ẹ̀ni tó dára jù lórí scoreboard.
- **Open rubric yí iwúrí padà (nínú ọ̀ràn iwádíí wọn).** Wọn dán ọ̀nà ríṣẹ̀rùn láti dín hallucination kù wò (jẹ́ kí àwòṣe dáhùn lẹ́ẹ̀mejì, kí ó sì kò láti dáhùn bí èsì méjèèjì kò bá mu). Lábẹ́ ipéye tó wópò, ọ̀nà nàà dín àṣìṣe kù *sùgbón ó tún dín ipéye kù* — nítorí nàà metric nàà ní dẹkun lílò rẹ̀. Lábẹ́ open rubric, ọ̀nà kan nàà borí fún gbogbo àwòṣe méréin lábẹ́ oríṣíríṣi ìjìyà; GPT-5-mini — tí raw ipéye fi ìjìyà jẹ́ nítorí ìkò láti dáhùn nígbà àní-àní — sì borí o4-mini nígbà tí a sọ ọ̀nà fífúnni ní máàkì ní gbangba ($n = 4,326$ ìbèèrè fún àwòṣe kòòkan).

Ohun tí èyí ṣeé ṣe kí ó túmò sí

Dídín hallucination kù kí í ṣe ní pàtàkì nípa dídá ìdánwò mii tó jẹ́ àkànṣe sí hallucination sílẹ̀. Ó jẹ́ nípa yíyí bí àwọn benchmark pàtàkì ṣe ní fún àní-àní ní máàkì, kí jíjẹwọ́ “Mi ò mò” má bàa tún gba ìjìyà. Títí scoreboard yóò fi yí padà, dídín hallucination kù yóò máa ná àwọn àwòṣe ní máàkì ipéye, a ó sì máa dẹkun rẹ̀ — idí nìyẹn tí àwọn ò̀nkòwé fi pe ìṣòro nàà ní “socio-technical”: apá kan jẹ́ metric tó dára sí i, apá kejì sì jẹ́ mímú àwọn leaderboard tó ní ipa gba á.

Ohun tí èyí kò fi hàn

- **Kò fi hàn pé open rubric ní tún hallucination ṣe ní ayé gidi.** Ìdánwò tó ṣe àtìlẹ́yìn fún un jẹ́ ọ̀ràn iwádíí kékeré tí a mọ́mọ́ ṣe **láìsì ìṣàkóso** — àwòṣe méréin ní ètò default, ọ̀nà ìdínkù kan, ìdánwò factual-QA kan — láti fi *ìyípadà iwúrí* hàn, kí í ṣe láti to àwòṣe ní ipò tàbí fi ipa gbogboogbo múlẹ̀.
- **Kò sọ pé fífúnni ní máàkì ni idí kan soṣo.** Àṣìṣe nínú dátà ìkòni, ìṣòro tó nira gidi, àti prompt tí àwòṣe kò

mò sì jẹ orísun mǐràn.

- Kò **ṣe àtílẹ̀yìn fún ọ̀rọ̀ olókíkí pé hallucination kò ṣeé yẹ**. Àwọn òhàkòwé n jiyàn ní òdikejì: ètò kan tó dáhùn ìbèèrè tó ṣeé ṣàyèwò nìkan, tó sì sọ “Mi ò mò” nígbà mǐràn, kò ní hallucinate lác.
- Kò **mú ààlà ìsàlẹ̀ pretraining tán** — ó ṣàlàyé rẹ, ó sì fi ààlà sí i; ààlà nàà kan àṣiṣe òtítọ́ tí a sọ pẹ̀lú idánílójú, kì í ṣe gbogbo ihùwàsí àwòṣe.
- Kò **fi hàn pé open rubric tó fúnra rẹ**. Ó n yí ohun tí idánwò n san èrè fún padà; kì í ṣe arópò retrieval, lílo irinṣẹ́, tàbí àwòṣe tí calibration rẹ dára sí i.

Báwo ni ẹ̀rí nàà ṣe lágbára tó?

- Ìpilẹ̀ nàà jẹ **mathematiki** — ààlà ìsàlẹ̀ àṣe, kì í ṣe ọ̀ṣùwòn. Gégé bí àrìyànjìyàn àbá-ẹ̀kọ́, ó fẹ̀ṣẹ̀ múlẹ̀ lábẹ̀ àwọn àfojúsùn tirẹ.
- Ó dúró lórí **àwòṣe ìṣòro tí a mòdómò mú rọ̀rùn**; àwọn òhàkòwé fúnra wọn tọka sí “false trichotomy” tí pípá gbogbo èsì sí tó tó, tó ṣàṣiṣe, tàbí “Mi ò mò”, àti ipò “òótọ́ tí kò ní àpẹ̀rẹ” tí wọn lò fún ààlà tó mò jù.
- Ìwádíí benchmark jẹ **àpẹ̀rẹ̀ kékeré tí a yàn** — idánwò olókíkí mewaá, kì í ṣe audit gbogbo rẹ.
- Case ìwádíí nàà jẹ **gidi sùgbón ó ní ààlà**: frontier àwòṣe méréin, ọ̀nà idínkù kan, SimpleQA nikan, ètò default, tí a sì sọ ní kedere pé “kì í ṣe idánwò tó ní ìṣàkóso.” Ó jẹ proof of concept fún àrìyànjìyàn iwúrí, kì í ṣe èsì benchmark.
- Ó yẹ ká sọ ibi tí iwòye nàà ti wá: mètá nínú òhàkòwé méréin n ṣiṣẹ̀ tàbí ti ṣiṣẹ̀ ní **OpenAI**, iwé nàà sì n jiyàn pé ẹ̀ka nàà yẹ kí ó yí ọ̀nà tó fi n dán àwòṣe wò padà. Èyí jẹ idúró tí a ṣàlàyé dáadàa láti ọ̀dọ̀ eni tó ní ipa nínú ọ̀rọ̀ nàà, kì í ṣe àyèwò ọ̀de aláìṣojúsàájú — ohun tó yẹ ká wọn, kì í ṣe ká kò. (Ohun tó dára ni pé iwé nàà fi àrìwísí kan àwọn àwòṣe tiwọn, o4-mini àti GPT-5-mini, gégé bí ó ṣe kan àwọn mǐ.)

Ìdí tí ó fi ṣe pàtàkì

Ó tún ọ̀nà tí a fi n wo ìṣòro tí a ti polówó rẹ gan-an ṣe. Wọn sàbà máa n ta “hallucination” bí àbùkù àjẹ̀jì tàbí odi tí kò ṣeé kojá; iwé yí mú un jẹ ohun lasán tí a sì dá sílẹ̀ ní apá kan — ààlà ìsàlẹ̀ statistiki tí a lè lóye, tó jókòò lórí iwúrí tí àwa fúnra wa yàn. Ẹ̀kọ́ tó gbòdòrò jẹ diẹ̀dìẹ̀, ó sì wúlò jù: ìlọ̀síwájú sí i nínú reliability lè sinmi lórí *ohun tí a n díwọn* gégé bí ó ti sinmi lórí ohun tí a n kọ.

Àkótán mímó

Àwọn èsì èké tí èdè àwòṣe n sọ pẹ̀lú idánílójú wá láti ibi méjì. Àkókọ́ jẹ statistiki: nígbà tí òtítọ́ kan kò ní àpẹ̀rẹ̀ láti kọ, àwòṣe tí a kọ láti fara wé èdè yóò ṣàṣiṣe nígbà mǐràn, a sì lè fojú díwọn ààlà ìsàlẹ̀ yẹn (fún àpẹ̀rẹ̀, láti iye òtítọ́ tó fara hàn lẹ̀ẹ̀kan soṣo nínú ìkọ̀nì). Èlẹ̀ẹ̀kejì ni iwúrí: fẹ̀rẹ̀ gbogbo benchmark tí a fi n tó àwòṣe ní ipò n fún “Mi ò mò” ní mǎàkì kan nàà pẹ̀lú èsì àṣiṣe, nítorí nàà iméfò máa n borí — títi dé ibi tí àwòṣe tó ṣàṣiṣe ní idáméta-méréin àkókò lè ju èyí tó jẹ olóótítọ́ sí i tó sì kọ láti dáhùn lọ. Àbá àwọn òhàkòwé kì í ṣe idánwò hallucination mǐ, bí kò ṣe “open rubric”: sọ ọ̀nà fífúnni ní mǎàkì nínú ìbèèrè. Nínú ọ̀ràn ìwádíí lórí frontier àwòṣe méréin, èyí yí iwúrí

padà kí ọ̀nà tó ní dín hallucination kù gba èrè dípò ìjìyà. Ó jẹ̀ iwé theory-pẹ̀lú-ìwádíí, pẹ̀lú ìdánwò kékeré tí a sọ pé kò ní ìṣàkóso, tí a ṣe peer review rẹ̀ nínú *Nature*; àtúnṣe náà ní ìlérí sùgbón a kò tì fi hàn pé ó ní ṣìṣẹ̀ ní ìwọn òlák, àwọn òhàkòwé sì jìyàn pé hallucination kì í ṣe àdìitú tàbí ohun tí kò ṣeé yẹ pátápátá.

Àyèwò láìsí ọ̀rò asán

Ohun tí iwé náà fi hàn: Ààlà ìsàlẹ̀ mathematiki tí ó mú kí hallucination diẹ̀ jẹ̀ dandan nígbà pretraining (ó kéré tán “iye singleton” fún òtítọ̀ tí kò ní àpẹrẹ̀); ìwádíí tó rí i pé ọ̀pọ̀ benchmark aṣáájú kò fún “Mi ò mò” ní máàkì; àti ọ̀ràn ìwádíí àwòṣe méréń níbi tí síṣọ̀ ọ̀nà fífúnni ní máàkì nínú prompt (“open rubric”) ti mú kí ọ̀nà dídín hallucination kù borí, níbi tí ìpéye lasán ti fi ìjìyà jẹ̀ ẹ̀.

Ohun tó ṣeé gbà sùgbón tí a kò fídí rẹ̀ múlẹ̀: Pé open rubric, bí a bá fi sínú àwọn benchmark pàtàkì, yóò dín hallucination nínú àwòṣe tí a fi sílò kù ní ọ̀nà tó ní ìtumọ̀. Ìdánwò tó ṣe àtílẹ̀yìn fún un kéré, a sì sọ ní kedere pé kò ní ìṣàkóso.

Ohun tí kò fi hàn: Pé hallucination kò ṣeé yẹ (ó jìyàn ní òdìkẹ̀jì); pé fífúnni ní máàkì ni ìdí kan ṣoṣo; pé a lè pa hallucination rẹ̀ ráúrú; tàbí pé ọ̀ràn ìwádíí náà ní to àwòṣe méréń sí ipò.

Àwọn ààlà pàtàkì: Àwòṣe tó mòdómò mú ìṣòro rọ̀rùn sí tó tọ̀/tó ṣàṣìṣe/“Mi ò mò” (àwọn òhàkòwé pè é ní “false trichotomy”); ìwádíí benchmark kékeré tí a yàn (ìdánwò mewaá); ọ̀ràn ìwádíí tí kò ní ìṣàkóso (àwòṣe méréń, ọ̀nà ìdínkù kan, ìdánwò kan, ètò default); àti àrìyànjiyàn tí OpenAI darí nípa bí ẹ̀ka náà ṣe yẹ kí ó máa dán àwòṣe wò.

Ìgbẹ̀kẹ̀lé mélòó ni olùkà gbogboogbo yẹ kí ó ní? Ìgbẹ̀kẹ̀lé gíga pé hallucination kì í ṣe àdìitú tàbí ohun tí kò ṣeé yẹ pátápátá, àti pé benchmark pàtàkì ní san èrè fún ìmẹ̀fò ní bá yí. Ìgbẹ̀kẹ̀lé àárín pé àtúnṣe tí a dábàá yóò ran lówó — proof of concept gidi wà bá yí, sùgbón kò tì sí ifihàn pé ó ní ṣìṣẹ̀ káàkiri àti ní ìwọn òlák.

Àwọn orísun

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>