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 ìdánilójú kì í ṣe àbùkù àdìítú: àwọn kan jẹ́ èsì statistiki ti ìkọ́ni, wọ́n sì ń dúró nítorí pé benchmark pàtàkì ń san èrè fún ìméfò pẹ̀lú ìdánilójú ju "Mi ò mọ̀" tó jẹ́ òtítọ́ lọ. Case study lórí frontier model mérin 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 · 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 model fi ń méfò, àti ìdí tí a fi kọ́ wọn láti ṣe bẹ́ẹ̀

Béèrè lọ́wọ́ large language model nípa ọjọ́ ìbí àjèjì kan, ó lè dáhùn “March 7” pẹ̀lú ìdánilójú bí ẹni tó ń kà á lórí káàdì — kí ó sì ṣàṣìṣe lẹ́ẹ̀mẹta ọ̀tọ̀ọ̀tọ̀ pẹ̀lú ọjọ́ méta tó yàtọ̀. Àwọn òǹkọ̀wé fúnni ní irú àpẹẹrẹ yìí gan-an: wọ́n béèrè ìbéèrè òtítọ́ lasán lọ́wọ́ àwọn model aṣáájú — ọjọ́ ìbí ènìyàn kan, tàbí ìtumọ̀ acronym kan tí a kò mọ̀ púpọ̀ — ọkọọkan wọn sì dá èsì míì sílẹ̀ pẹ̀lú ìdánilójú, tí kò sí èyíkéyìí tó tọ́. Ọ̀rọ̀ tí ilé-iṣẹ́ ń lò fún èyí ni *hallucination*, tó mú un dà bí àṣìṣe nínú ohun tí a rí. Ìgbésẹ̀ àkọ́kọ́ ìwé náà ni láti yọ àdìítú kúrò nínú rẹ̀.

Bẹ̀rẹ̀ pẹ̀lú bí a ṣe kọ́ model. Ní ìpele ìkọ́ni àkọ́kọ́ tó sì tóbi jù lọ, ó kẹ́kọ̀ọ́, ní kókó, bí èdè tó ń ṣàn dáadáa ṣe rí nípa kíka ọ̀pọ̀lọpọ̀ ọ̀rọ̀. Wá mú òtítọ́ kan tí kò ní àpẹẹrẹ lẹ́yìn rẹ̀ — ọjọ́ ìbí ènìyàn pàtó kan. Bí ọjọ́ náà bá fara hàn nínú ọ̀rọ̀ ìkọ́ni lẹ́ẹ̀kan, tàbí kò fara hàn rárá, kò sí ohun tí akẹ́kọ̀ọ́-àpẹẹrẹ lè dì mú: láti ojú ìwòye model, èsì náà jẹ́ ohun tí kò ní ìlànà. Àwọn òǹkọ̀wé sọ èyí di pípé nípa lílo èrò àtijọ́ kan (ti Alan Turing, láti inú ìṣòro míì): bí ọjọ́ ìbí kan nínú márùn-ún bá fara hàn lẹ́ẹ̀kan ṣoṣo nínú dátà, a yẹ kí a retí pé model yóò ṣàṣìṣe ó kéré tán lórí ọ̀kan nínú márùn-ún — kì í ṣe nítorí pé ó bàjẹ́, ṣùgbọ́n nítorí pé kò sí ohun kan níbẹ̀ láti kọ́. (Pẹ̀lú ìrònú kan náà, model fẹ́rẹ́é má ṣàṣìṣe lórí olú-ìlú orílẹ̀-èdè: wọ́n máa ń fara hàn ní gbogbo ìgbà.) Wọ́n jiyàn pẹ̀lú ìṣọ́ra pé yíya gbólóhùn òtítọ́ sọ́tọ̀ kúrò nínú èké tó ṣeé gbà jẹ́ ìṣòro tó nira fúnra rẹ̀, àti pé pípèsè gbólóhùn òtítọ́ *nìkan* ó kéré tán le tó ìṣòro yẹn. Ìpele àṣìṣe kan wà nínú ètò náà láti ìbẹ̀rẹ̀.

Èrò tó wà lábẹ́ rẹ̀: bí a ṣe ń ka ohun tí a kò tíì rí

Èyí dúró lórí èrò ọlọ́gbọ́n gidi kan — tó dàgbà ju language model lọ, tó sì yẹ kí a mọ̀ dáadáa.

Bẹ̀rẹ̀ pẹ̀lú àpò bọ́ọ̀lù aláwọ̀. O kò mọ iye àwọ̀ tó wà nínú rẹ̀. O fa 100 jáde, ọkọ̀ọ̀kan lẹ́ẹ̀ẹ̀sẹ̀, o sì ka wọn: pupa 40, búlúù 25, ewé 15, yẹ́lò 5, aláwọ̀ elése àlùkò 3, ọsàn 2 — lẹ́yìn náà **àwọ̀ mẹ́wàá ọ̀tọ̀ọ̀tọ̀ tí ọkọọkan fara hàn lẹ́ẹ̀kan ṣoṣo.**

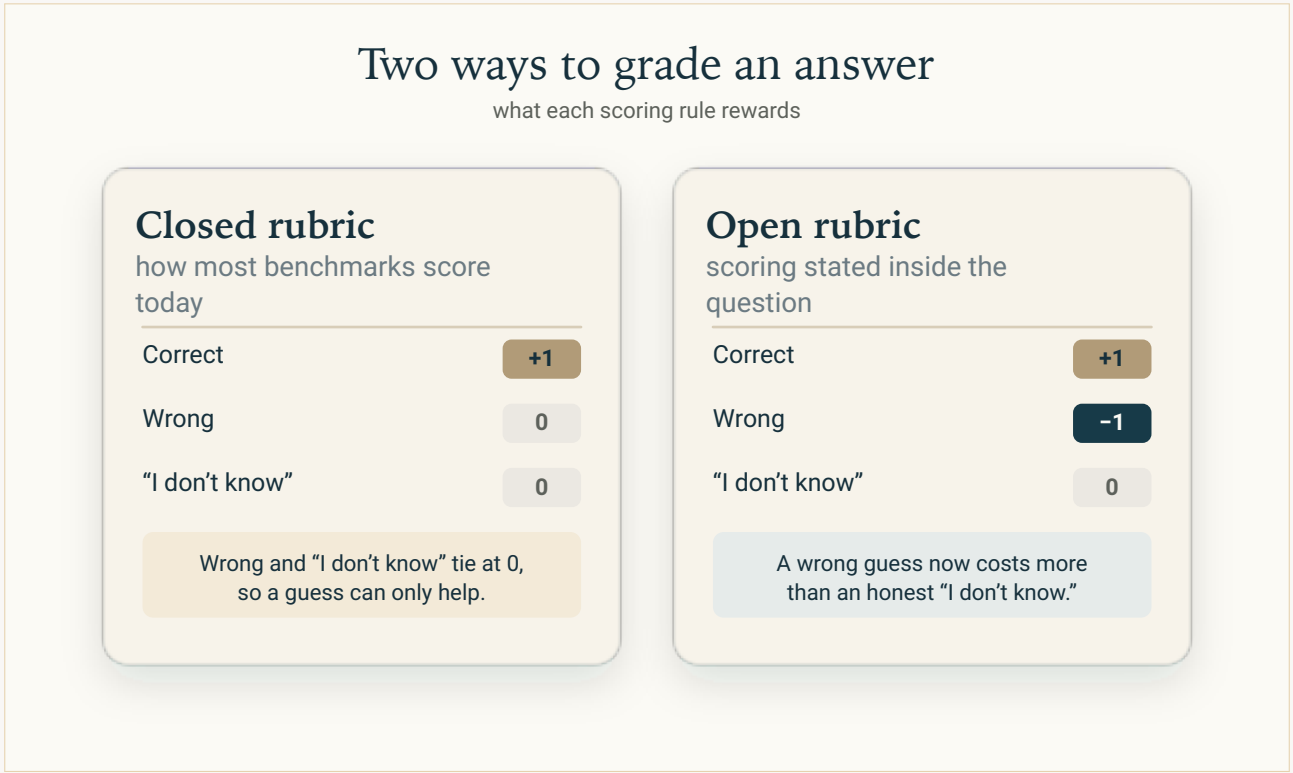
Wá wo ìbéèrè tí Turing dojú kọ ní tòótọ́, lórí ìṣòro tó yàtọ̀: *kí ni àǹfààní pé bọ́ọ̀lù tó kàn jẹ́ àwọ̀ tí o kò tíì rí rárá?* O kò lè ka ohun tí o kò tíì fa jáde — ṣùgbọ́n o **lè** ka àwọn àwọ̀ tí o ti rí lẹ́ẹ̀kan ṣoṣo, àwọn “singleton.” Ọgbọ́n náà, tí a ń pè ní **Good–Turing estimation**, ni pé ìpín àwọn ìfà rẹ tó jẹ́ singleton ń fojú díwọ̀n probability tó ṣì fara pamọ́ sínú àwọn àwọ̀ tí o kò tíì rí. Mẹ́wàá nínú ọgọ́rùn-ún ìfà rẹ jẹ́ àwọ̀ tí o rí lẹ́ẹ̀kan ṣoṣo, nítorí náà àǹfààní pé bọ́ọ̀lù tó kàn jẹ́ àwọ̀ tuntun pátápátá jẹ́ nǹkan bí **10 / 100 = 10%.**

Àwọn àwọ̀ tí a rí lẹ́ẹ̀kan náà kì í ṣe *àṣìṣe*. Wọ́n jẹ́ òṣùwọ̀n àìmọ̀ tìrẹ: bí ọ̀pọ̀ àwọ̀ bá fara hàn lẹ́ẹ̀kan, àpẹẹrẹ náà ń sọ fún ọ pé ayé ní púpọ̀ sí i tí o kò tíì fa jáde.

Wá fi ọjọ́ ìbí rọ́pò àwọ̀, kí o sì fi ọ̀rọ̀ ìkọ́ni model rọ́pò àpò náà. Ká sọ pé, nínú àwọn ọjọ́ ìbí tí ó rí, **ọ̀kan nínú márùn-ún fara hàn lẹ́ẹ̀kan ṣoṣo.** Ọgbọ́n kan náà: nǹkan bí ìdá márùn-ún probability wà nínú àwọn ọjọ́ ìbí tí model kò rí ní tòótọ́ — ọjọ́ ìbí kò sì ní àpẹẹrẹ tí a lè gbára lé (o kò lè *ṣírò* ọjọ́ ìbí ẹnì kan). Nítorí náà, ọjọ́ tí a rí lẹ́ẹ̀kan, tàbí tí a kò rí rárá, dà bí owó tí model kò lè fi ìwọ̀n sí, yóò sì ṣàṣìṣe lórí nǹkan bí **ọ̀kan nínú márùn-ún** wọn. Kò sí ọgbọ́n tó lè tún èyí ṣe: kò sí ohun níbẹ̀ láti kọ́.

> Èyí ni gbogbo àríyànjiyàn náà ní kúkúrú: iye singleton ń díwọ̀n iye ayé tí a kò lè kọ́ *láti inú dátà yìí*, èyí sì di **ààlà ìsàlẹ̀** fún àṣìṣe. Ó tún jẹ́ ìdí tí model fẹ́rẹ́é má ṣe ṣàṣìṣe olú-ìlú kan — *Paris* fara hàn ní gbogbo ìgbà, iye singleton rẹ̀ sún mọ́ òdo, nítorí náà ọ̀pọ̀lọpọ̀ ohun ló wà láti kọ́.

Èyí ṣàlàyé ibi tí hallucination ti wá. Kò ṣàlàyé ìdí tí wọ́n fi *yè* — ìdí tí àwọn model, lẹ́yìn gbogbo ìkọ́ni tó tẹ̀lé tí a ṣe láti mú wọn wúlò kí wọ́n sì jẹ́ olóòtítọ́, ṣì ń ṣe bí ẹni pé wọ́n mọ̀ dípò kí wọ́n jẹ́wọ́ àní-àní. Níhìn-ín, àfiwé ìwé náà fẹ́rẹ́é jẹ́ ohun tó ń kó ènìyàn ní ìtìjú. Foju inú wo akẹ́kọ̀ọ́ nínú ìdánwò tí kò mọ èsì. Bí ààyè òfo bá gba òdo, tí ìméfò sì lè gba ọ̀kan, ìgbésẹ̀ tó ń mú máàkì pọ̀ jù ni láti méfò — pẹ̀lú ìdánilójú àti kíkún, kì í ṣe “Mi ò dájú.” Àwọn akẹ́kọ̀ọ́ ń kọ́ èyí. Ó hàn pé àwọn model náà ń kọ́ ọ — nítorí pé báyìí náà ni a ṣe ń fún wọn ní máàkì. Àwọn òǹkọ̀wé ṣàyẹ̀wò àwọn benchmark tí ẹ̀ka náà ń díje lórí gan-an, àwọn leaderboard tí a ń ṣètò model láti gòkè, wọ́n sì rí i pé fẹ́rẹ́é gbogbo wọn ń fún “Mi ò mọ̀” ní máàkì kan náà pẹ̀lú èsì àṣìṣe: òdo. Lábẹ́ òfin yẹn, model tó máa ń méfò ní gbogbo ìgbà yóò borí model tó jọ ọ́ ní gbogbo ọ̀nà míì ṣùgbọ́n tó ń sọ àní-àní rẹ̀ ní òtítọ́. Ní ìtumọ̀ tó fẹ́rẹ́é jẹ́ tàárà, a ń fi máàkì kọ́ wọn láti ṣe é.



Àwọn benchmark lè mú ìméfò jẹ́ ohun tó bójú mu. Lábẹ́ ọ̀nà fífúnni ní máàkì tí ọ̀pọ̀ benchmark ń lò (òsì), èsì àṣìṣe àti “Mi ò mọ̀” tó jẹ́ òtítọ́ ń gba òdo — nítorí náà ìméfò lè ṣèrànwọ́ nìkan. Ṣọ àwọn ọ̀fin náà nínú ìbéèrè, pẹ̀lú ìjìyà fún àṣìṣe (ọ̀tún), 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.

Apá yìí ló yẹ ká pa mọ́, nítorí pé ó lòdì sí àkọlé tí a sábà rí. Wọ́n sábà máa ń ta hallucination gẹ́gẹ́ bí ààlà tí kò ṣeé yẹ̀, tó fẹ́rẹ́é jẹ́ ohun ìjìnlẹ̀ ti technology. Ìwé náà tako ìsọ̀rọ̀ méjèèjì. Ààlà

ìsàlẹ̀ pretraining kì í ṣe àdìítú — àṣìṣe statistiki lasán ni, irú èyí tí machine learning ti mọ̀ fún ọ̀pọ̀ ọdún. Ìdúró rẹ̀ pé náà kò sì ṣe dandan — ní apá kan, ìwúrí tí àwa fúnra wa kọ́ ni, a sì lè yí i padà. Èrò kan tó kàn kọ́ láti dáhùn nígbà tí kò dájú kò ní hallucinate rárá; ìdí tí àwọn model tí a ń lò kò fi ṣe bẹ́ẹ̀ ni pé scoreboard wa ń fi ìjìyà jẹ ìkọ́ láti dáhùn.

Ohun tí àwọn òǹkọ̀wé ṣe

Ìwé náà ní apá mẹ́ta. Àkọ́kọ́, àríyànjiyàn mathematiki pé hallucination díẹ̀ jẹ́ dandan ní statistiki nígbà pretraining, nípa fífi hàn pé “pèsè ọ̀rọ̀ tó fẹsẹ̀ múlẹ̀ nìkan” ó kéré tán le tó ìṣòro classification alakomeji “ṣé gbólóhùn yìí fẹsẹ̀ múlẹ̀?” Ẹlẹ́ẹ̀kejì, àríyànjiyàn — tí ìwádìí benchmark olókìkí mẹ́wàá ṣe àtìlẹ́yìn fún — pé àwọn metric accuracy tó wọ́pọ̀ ń san èrè fún ìméfò ju ìkọ́ láti dáhùn lọ. Ẹlẹ́ẹ̀kẹta, àtúnṣe tí wọ́n dábàá àti case study tó dán an wò: ìdánwò **open-rubric**, níbi tí a ti sọ òṣì fífúnni ní máàkì nínú ìbéèrè fúnra rẹ̀ (fún àpẹẹrẹ, “èsì tó tọ́ gba 1, èyí tó ṣàṣìṣe gba −1, nítorí náà má dáhùn bí ìdánilójú rẹ kò bá tó 50%”), kí model lè mọ ìgbà tí a ń san èrè fún òtítọ́. Wọ́n dán èyí wò lórí frontier model mẹ́rin — Gemini 3 Pro ti Google, GPT-5 ti OpenAI, Grok 4 ti xAI àti Claude Opus 4.5 ti Anthropic — ní lílo àwọn ìbéèrè òtítọ́ 4,326 ti SimpleQA. Wọ́n sọ ní kedere pé case study náà jẹ́ àpẹẹrẹ, “kì í ṣe ìdánwò tó ní ìṣàkóso láàárín model” (èrò default, kò sí tuning, kò sí ìṣọ̀kan owó).

Ohun tí wọ́n rí

- **Pretraining ń fipá mú àṣìṣe díẹ̀.** Iye tí model fi ń sọ èké pẹ̀lú ìdánilójú ní àálà ìsàlẹ̀ tó jẹ́ (ní àfojúsùn, ìlọ́po méjì) iye àṣìṣe classifier “ṣé gbólóhùn yìí fẹsẹ̀ múlẹ̀?” tó dára jù lọ tí a kọ́ láti inú rẹ̀. Fún àwọn òtítọ́ tí kò ní àpẹẹrẹ tí a lè kọ́, àálà ìsàlẹ̀ náà ó kéré tán jẹ́ *iye singleton* — ìpín àwọn òtítọ́ tí ó fara hàn lẹ́ẹ̀kan ṣoṣo nínú ìkọ́ni. Hallucination díẹ̀ kò ṣeé yẹ̀ kódà pẹ̀lú dátà tó mọ́ pátápátá.
- **Fífúnni ní máàkì ń san èrè fún ìméfò — ní kedere.** Lábẹ́ fífúnni ní máàkì tó ń sọ pé ó tọ́ tàbí ó ṣàṣìṣe, àìkọ́ láti dáhùn láé ni strategy tó dára jù, ìwádìí àwọn òǹkọ̀wé sì rí i pé ọ̀pọ̀ jù lọ benchmark olókìkí ń ka “Mi ò mọ̀” sí àṣìṣe lasán. Àpẹẹrẹ tó hàn gbangba láti ọ̀dọ̀ wọn: lórí ìdánwò SimpleQA, raw accuracy díẹ̀ *ṣe ojú rere sí* o4-mini ti OpenAI — tó dáhùn fẹ́rẹ́é ohun gbogbo tí ó sì ṣàṣìṣe ju ìdámẹ́ta-mẹ́rin lọ — ju GPT-5-mini, tó ṣe àṣìṣe díẹ̀ púpọ̀ nítorí pé ó kọ́ láti dáhùn nígbà àní-àní. Model tó ń ṣe ewu jù ló dà bí ẹni tó dára jù lórí scoreboard.
- **Open rubric yí ìwúrí padà (nínú case study wọn).** Wọ́n dán ọ̀nà rírọ̀rùn láti dín hallucination kù wò (jẹ́ kí model dáhùn lẹ́ẹ̀mejì, kí ó sì kọ́ láti dáhùn bí èsì méjèèjì kò bá mu). Lábẹ́ accuracy tó wọ́pọ̀, ọ̀nà náà dín àṣìṣe kù *ṣùgbọ́n ó tún dín accuracy kù* — nítorí náà metric náà ń dẹ́kun lílò rẹ̀. Lábẹ́ open rubric, ọ̀nà kan náà borí fún gbogbo model mẹ́rin lábẹ́ orísìírísi ìjìyà; GPT-5-mini — tí raw accuracy fi ìjìyà jẹ nítorí ìkọ́ láti dáhùn nígbà àní-àní — sì borí o4-mini nígbà tí a sọ ọ̀nà fífúnni ní máàkì ní gbangba (n = 4,326 ìbéèrè fún model kọ̀ọ̀kan).

Ohun tí èyí ṣeé ṣe kí ó túmọ̀ sí

Dídín hallucination kù kì í ṣe ní pàtàkì nípa dídá ìdánwò míì tó jẹ́ àkànṣe sí hallucination sílẹ̀. Ó jẹ́ nípa yíyí bí àwọn benchmark pàtàkì ṣe ń fún àní-àní ní máàkì, kí jíjẹ́wọ́ “Mi ò mọ̀” má bàa tún gba ìjìyà. Títí scoreboard yóò fi yí padà, dídín hallucination kù yóò máa ná àwọn model ní máàkì accuracy, a ó sì máa dẹ́kun rẹ̀ — ìdí nìyẹn tí àwọn òǹkọ̀wé fi pe ìṣòro náà 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 ń tún hallucination ṣe ní ayé gidi**. Ìdánwò tó ṣe àtìlẹ́yìn fún un jẹ́ case study kékeré tí a mọ̀ọ́mọ̀ ṣe **láìsí ìṣàkóso** — model mẹ́rin ní ètò default, ọ̀nà ìdínkù kan, ìdánwò factual-QA kan — láti fi *ìyípadà ìwúrí* hàn, kì í ṣe láti to model ní ipò tàbí fi ipa gbogboogbo múlẹ̀.
- Kò **sọ pé fífúnni ní máàkì ni *ìdí kan ṣoṣo***. Àṣìṣe nínú dátà ìkọ́ni, ìṣòro tó nira gidi, àti prompt tí model kò mọ̀ ṣì jẹ́ orísun mìíràn.
- Kò **ṣe àtìlẹ́yìn fún ọ̀rọ̀ olókìkí pé hallucination kò ṣeé yẹ̀**. Àwọn òǹkọ̀wé ń jiyàn ní òdìkejì: ètò kan tó dáhùn ìbéèrè tó ṣeé ṣàyẹ̀wò nìkan, tó sì sọ “Mi ò mọ̀” nígbà mìíràn, kò ní hallucinate láé.
- Kò **mú ààlà ìsàlẹ̀ pretraining tán** — ó ṣàlàyé rẹ̀, ó sì fi ààlà sí i; ààlà náà kan àṣìṣe òtítọ́ tí a sọ pẹ̀lú ìdánilójú, kì í ṣe gbogbo ìhùwàsí model.
- Kò **fi hàn pé open rubric *tó* fúnra rẹ̀**. Ó ń yí ohun tí ìdánwò ń san èrè fún padà; kì í ṣe arọ́pò retrieval, lílo tool, tàbí model tí calibration rẹ̀ dára sí i.

Báwo ni ẹ̀rí náà ṣe lágbára tó?

- Ìpìlẹ̀ náà jẹ́ **mathematiki** — ààlà ìsàlẹ̀ àṣẹ, kì í ṣe òṣùwọ̀n. Gégé bí àríyànjiyàn theory, ó fẹsẹ̀ múlẹ̀ lábẹ́ àwọn àfojúsùn tirẹ̀.
- Ó dúró lórí **àwòṣe ìṣòro tí a mọ̀ọ́mọ̀ mú rọrùn**; àwọn òǹkọ̀wé fúnra wọn tọ́ka sí “false trichotomy” ti pípa gbogbo èsì sí tó tọ́, tó ṣàṣìṣe, tàbí “Mi ò mọ̀”, àti ipò “òótọ́ tí kò ní àpẹẹrẹ” tí wọ́n lò fún ààlà tó mọ́ jù.
- Ìwádìí benchmark jẹ́ **àpẹẹrẹ kékeré tí a yàn** — ìdánwò olókìkí mẹ́wàá, kì í ṣe audit gbogbo rẹ̀.
- Case study náà jẹ́ **gidi ṣùgbọ́n ó ní ààlà**: frontier model mẹ́rin, ọ̀nà ìdínkù kan, SimpleQA nìkan, ètò default, tí a sì sọ ní kedere pé “kì í ṣe ìdánwò tó ní ìṣàkóso.” Ó jẹ́ proof of concept fún àríyànjiyàn ìwúrí, kì í ṣe èsì benchmark.
- Ó yẹ ká sọ ibi tí ìwòye náà ti wá: mẹ́ta nínú òǹkọ̀wé mẹ́rin ń ṣiṣẹ́ tàbí ti ṣiṣẹ́ ní **OpenAI**, ìwé náà sì ń jiyàn pé ẹ̀ka náà yẹ kí ó yí ọ̀nà tó fi ń dán model wò padà. Èyí jẹ́ ìdúró tí a ṣàlàyé dáadáa láti ọ̀dọ̀ ẹni tó ní ipa nínú ọ̀rọ̀ náà, kì í ṣe àyẹ̀wò òde aláìṣojúsàájú — ohun tó yẹ ká

wọn, kì í ṣe ká kọ. (Ohun tó dára ni pé ìwé náà fi àríwísí kan àwọn model tiwọn, o4-mini àti GPT-5-mini, gẹ́gẹ́ bí ó ṣe kan àwọn míì.)

Ìdí tí ó fi ṣe pàtàkì

Ó tún ọ̀nà tí a fi ń wo ìṣòro tí a ti polówó rẹ̀ gan-an ṣe. Wọ́n sábà máa ń ta “hallucination” bí àbùkù àjèjì tàbí odi tí kò ṣeé kọjá; ìwé yìí mú un jẹ́ ohun lasán tí a sì dá sílẹ̀ ní apá kan — ààlà ìsàlẹ̀ statistiki tí a lè lóye, tó jókòó lórí ìwúrí tí àwa fúnra wa yàn. Ẹ̀kọ́ tó gbòòrò jẹ́ díẹ̀díẹ̀, ó sì wúlò jù: ìlọsíwájú sí i nínú reliability lè sinmi lórí *ohun tí a ń díwọ̀n* gẹ́gẹ́ bí ó ti sinmi lórí ohun tí a ń kọ́.

Àkótán mímọ́

Àwọn èsì èké tí language model ń sọ pẹ̀lú ìdánilójú wá láti ibi méjì. Àkọ́kọ́ jẹ́ statistiki: nígbà tí òtítọ́ kan kò ní àpẹẹrẹ láti kọ́, model tí a kọ́ láti fara wé èdè yóò ṣàṣìṣe nígbà mìíràn, a sì lè fojú díwọ̀n ààlà ìsàlẹ̀ yẹn (fún àpẹẹrẹ, láti iye òtítọ́ tó fara hàn lẹ́ẹ̀kan ṣoṣo nínú ìkọ́ni). Ẹlẹ́ẹ̀kejì ni ìwúrí: fẹ́rẹ́é gbogbo benchmark tí a fi ń to model ní ipò ń fún “Mi ò mọ̀” ní máàkì kan náà pẹ̀lú èsì àṣìṣe, nítorí náà ìméfò máa ń borí — títí dé ibi tí model tó ṣàṣìṣe ní ìdámẹ́ta-mẹ́rin àkókò lè ju èyí tó jẹ́ olóòtítọ́ sí i tó sì kọ̀ láti dáhùn lọ. Àbá àwọn òǹkọ̀wé kì í ṣe ìdánwò hallucination míì, bí kò ṣe “open rubric”: sọ ọ̀nà fífúnni ní máàkì nínú ìbéèrè. Nínú case study lórí frontier model mẹ́rin, èyí yí ìwúrí padà kí ọ̀nà tó ń dín hallucination kù gba èrè dípò ìjìyà. Ó jẹ́ ìwé theory-pẹ̀lú-ìwádìí, pẹ̀lú ìdánwò kékeré tí a sọ pé kò ní ìṣàkóso, tí a ṣe peer review rẹ̀ nínú *Nature*; àtúnṣe náà ní ìlérí ṣùgbọ́n a kò tíì fi hàn pé ó ń ṣiṣẹ́ ní ìwọ̀n ńlá, àwọn òǹkọ̀wé sì jiyàn pé hallucination kì í ṣe àdìítú tàbí ohun tí kò ṣeé yẹ̀ pátápátá.

Àyẹ̀wò láìsí ọ̀rọ̀ asán

Ohun tí ìwé náà fi hàn: Ààlà ìsàlẹ̀ mathematiki tí ó mú kí hallucination díẹ̀ jẹ́ dandan nígbà pretraining (ó kéré tán “iye singleton” fún òtítọ́ tí kò ní àpẹẹrẹ); ìwádìí tó rí i pé ọ̀pọ̀ benchmark aṣáájú kò fún “Mi ò mọ̀” ní máàkì; àti case study model mẹ́rin níbi tí sísọ ọ̀nà fífúnni ní máàkì nínú prompt (“open rubric”) ti mú kí ọ̀nà dídín hallucination kù borí, níbi tí accuracy lasán ti fi ìjìyà jẹ́ ẹ́.

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

Ohun tí kò fi hàn: Pé hallucination kò ṣeé yẹ̀ (ó jiyàn ní òdìkejì); pé fífúnni ní máàkì ni ìdí kan ṣoṣo; pé a lè pa hallucination rẹ́ ráúráú; tàbí pé case study náà ń to model mẹ́rin sí ipò.

Àwọn ààlà pàtàkì: Àwòṣe tó mọ̀ọ́mọ̀ mú ìṣòro rọrùn sí tó tọ́/tó ṣàṣìṣe/“Mi ò mọ̀” (àwọn òǹkọ̀wé pè é ní “false trichotomy”); ìwádìí benchmark kékeré tí a yàn (ìdánwò mẹ́wàá); case study tí kò ní ìṣàkóso (model mẹ́rin, ọ̀nà ìdínkù kan, ìdánwò kan, ètò default); àti àríyànjiyàn tí OpenAI darí nípa bí ẹ̀ka náà ṣe yẹ kí ó máa dán model wò.

Ìgbẹ́kẹ̀lé mélòó ni olùkà gbogboogbo yẹ kí ó ní? Ìgbẹ́kẹ̀lé gíga pé hallucination kì í ṣe àdìítú tàbí ohun tí kò ṣeé yẹ̀ pátápátá, àti pé benchmark pàtàkì ń san èrè fún ìméfò ní báyìí. Ìgbẹ́kẹ̀lé àárín pé àtúnṣe tí a dábàá yóò ran lọ́wọ́ — proof of concept gidi wà báyìí, ṣùgbọ́n kò tíì sí ìfihàn pé ó ń ṣiṣẹ́ káàkiri àti ní ìwọ̀n ńlá.

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