The Clean Paper

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

Af hverju líkön giska — og af hverju við kenndum þeim að gera það

Sjálfsöruggar rangfærslur tungumálalíkana eru að hluta tölfræðileg afleiðing þjálfunar og halda áfram vegna þess að mörg viðmið verðlauna örugga ágiskun en ekki heiðarlegt „ég veit það ekki“. Tilviksrannsókn á fjórum fremstu líkönum sýnir að þegar matsreglurnar eru sagðar skýrt í prompt-inu breytist þessi hvati. Þetta er bæði fræðileg neðri-marka röksemd og takmörkuð tilraun, ekki ein alhliða skýring á öllum ranghugmyndum.

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

Af hverju líkön giska — og af hverju við kenndum þeim að gera það

Spyrðu stórt mállíkan um afmælisdag ókunnugs einstaklings og það gæti svarað „7. mars“ með sömu ró og manneskja sem les dagsetninguna af korti — og haft rangt fyrir sér, þrisvar í röð, með þrjár mismunandi dagsetningar. Höfundarnir gefa nákvæmlega svona dæmi: fremstu líkön fá einfalda staðreyndaspurningu, til dæmis afmælisdag eða merkingu sjaldgæfrar skammstöfunar, og hvert þeirra býr til annað sjálfsöruggt svar sem er rangt. Í greininni er fyrsta skrefið að taka dulúðina úr því sem iðnaðurinn kallar *hallucination*.

Byrjum á því hvernig líkan er þjálfað. Í fyrsta og stærsta þjálfunarstiginu lærir það í raun hvernig reiprennandi texti lítur út með því að lesa gríðarlegt textamagn. Tökum svo staðreynd án mynsturs — afmælisdag ákveðinnar manneskju. Ef dagsetningin birtist einu sinni í þjálfunargögnunum, eða alls ekki, er ekkert almennt mynstur sem hægt er að læra. Frá sjónarhóli líkansins er svarið að mestu handahófskennt.

Höfundarnir gera þessa hugmynd formlega með eldri aðferð sem tengist Alan Turing: ef einn af hverjum fimm afmælisdagsetningum í gögnunum birtist aðeins einu sinni ætti að búast við að líkan missi að minnsta kosti af svipuðum hluta þeirra — ekki vegna þess að það sé bilað, heldur vegna þess að lítið eða ekkert var til að læra. Sama röksemd skýrir af hverju höfuðborgir landa eru miklu auðveldari: þær koma fyrir aftur og aftur. Höfundarnir sýna líka að aðgreina sanna staðhæfingu frá trúverðugri rangri er sjálft erfitt flokkunarvandamál, og að framleiða **eingöngu** sannar staðhæfingar getur ekki verið auðveldara. Það verður því ákveðið lágmark skekkju innbyggt í grunnþjálfunina.

Hugmyndin undir: hvernig á að telja það sem þú hefur ekki séð

Þetta byggir á sniðugri hugmynd sem er eldri en mállíkön.

Hugsum okkur poka af lituðum kúlum. Þú veist ekki hversu margir litir eru í pokanum. Þú dregur 100 kúlur og telur: 40 rauðar, 25 bláar, 15 grænar, 5 gular, 3 fjólubláar, 2 appelsínugular — og svo **tíu ólíka liti sem koma hver aðeins einu sinni**.

Spurningin sem Turing þurfti að glíma við í allt öðru samhengi var: *hversu líklegt er að næsta kúla verði litur sem þú hefur aldrei séð?* Þú getur ekki talið litina sem aldrei komu upp, en þú getur talið litina sem komu **einu sinni**. Í **Good–Turing-mati** gefur hlutfall þessara „einstæðinga“ mat á líkindamassanum sem liggur í því sem þú hefur ekki enn séð. Ef tíu af hundrað dráttum voru litir sem komu bara einu sinni er líkur á alveg nýjum lit í næsta drætti um **10 / 100 = 10%**.

Litirnir sem komu einu sinni eru ekki mistök. Þeir mæla eigin fáfræði úrtaksins: margir einstæðingar segja að heimurinn innihaldi meira sem úrtakið hefur ekki náð að sjá.

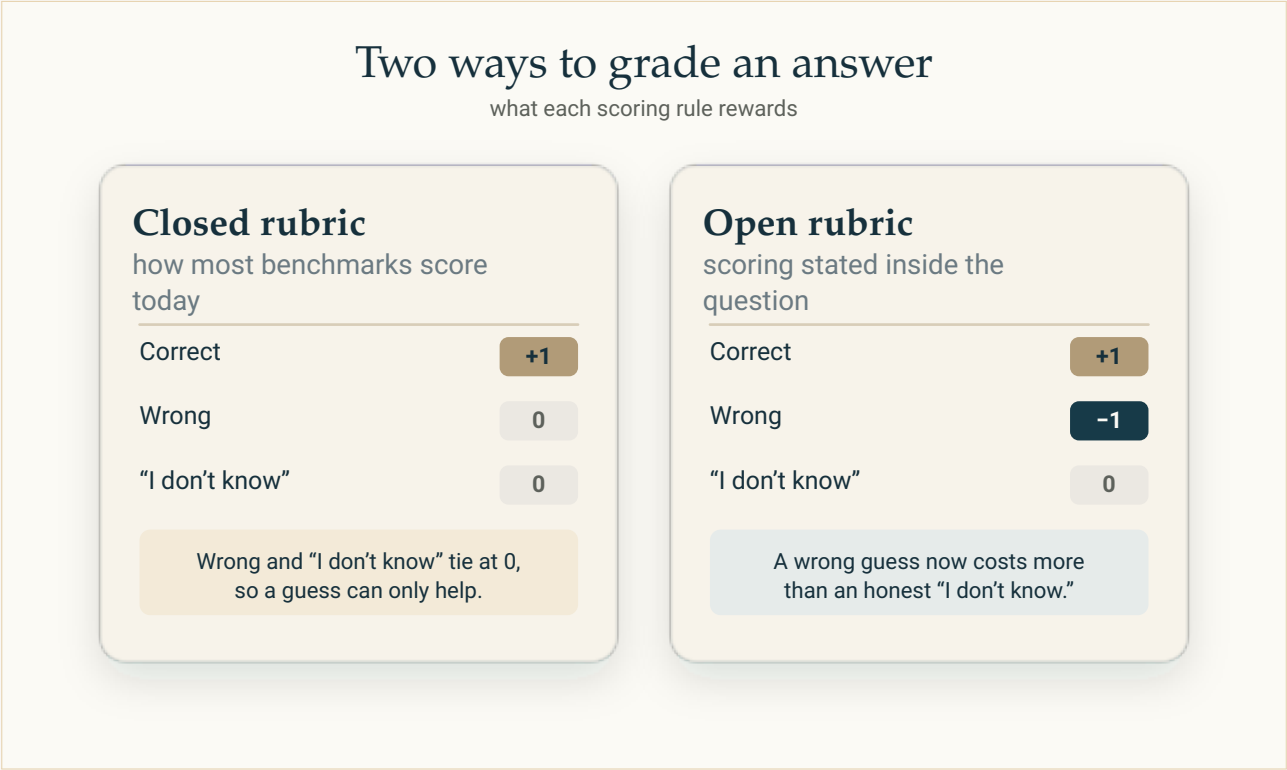
Skiptu nú litum út fyrir afmælisdaga og pokanum út fyrir þjálfunartexta líkans. Segjum að **einn af hverjum fimm** afmælisdögum sem líkanið sá hafi aðeins birst einu sinni. Sama aðferð segir að verulegur hluti líkinda liggi í dagsetningum sem líkanið hefur í reynd mjög lítil gögn um. Og afmælisdagur hefur ekkert almennt mynstur sem hægt er að reikna út. Dagsetning sem sást einu sinni eða aldrei er því staðreynd sem líkanið hefur veikan grunn til að spá fyrir um.

Þetta er meginröksemdin í smækkaðri mynd: tíðni einstæðinga mælir hversu mikið er

> ólærandi **úr þessum gögnum**, og það verður neðri mörk á skekkju. Þess vegna missir líkan sjaldan af höfuðborg eins og París: slík staðreynd kemur stöðugt fyrir og er því vel lærð.

Þetta skýrir hvaðan rangar staðreyndir geta komið. Það skýrir ekki hvers vegna þær **lifa af** seinni þjálfun sem á að gera líkön hjálplegri og heiðarlegri. Þar er líking greinarinnar næstum óþægilega góð. Hugsum okkur nemanda í prófi sem veit ekki svarið. Ef autt svar fær núll stig en ágiskun getur fengið eitt stig er stigahámarksstefnan að giska — ekki að segja „ég veit það ekki“. Nemendur læra þetta. Líkön gera það líka, vegna þess að við gefum þeim oft sömu hvata.

Höfundarnir skoðuðu þau viðmið sem sviðið raunverulega keppir á og þau stigatöflukerfi sem líkön eru fínstillt til að klífa. Í langflestum tilvikum fær „ég veit það ekki“ **sama skor og rangt svar: núll**. Undir slíkri reglu mun líkan sem giskar alltaf slá annars eins líkan sem segir heiðarlega frá óvissu. Að hluta til erum við því bókstaflega að **stiga líkön inn í giskhegðun**.



Viðmið geta gert ágiskun skynsamlega. Með algengu stigakerfi (vinstra megin) fá rangt svar og heiðarlegt „ég veit það ekki“ bæði núll, þannig að ágiskun getur aðeins hjálpað. Ef stigareglurnar eru settar inn í spurninguna og rangt svar fær refsingu (hægra megin) getur verið betra að sleppa svari þegar óvissan er mikil. Þetta breytir hvatanum; það leysir ekki sjálfkrafa öll rangsvör.

Original diagram — The Clean Paper CC BY 4.0

Þetta er sá hluti sem skiptir mestu máli vegna þess að hann gengur gegn algengri frásögn. „Hallucination“ er oft kynnt sem dularfull og óhjákvæmileg tæknileg takmörkun. Greinin hafnar báðum túlkunum. Skekkjan frá grunnþjálfun er ekki dularfull — hún er venjuleg tölfræðileg skekkja. Og viðvarandi giskhegðun er ekki algjörlega óhjákvæmileg — hún er að hluta hvati sem við sjálf hönnuðum og gætum breytt. Kerfi sem svaraði aðeins þegar það væri nógu öruggt og segði annars „ég veit það

ekki“ myndi ekki búa til sjálfsörugga staðreynd í hvert skipti; stigakerfin okkar gera hins vegar oft slíka varfærni óhagstæða.

Hvað gerðu höfundarnir?

Greinin hefur þrjá hluta.

Fyrst kemur stærðfræðileg röksemd um að einhver staðreyndaskekkja sé tölfræðilega óhjákvæmileg í grunnþjálfun. Höfundarnir sýna að vandamálið „framleiða aðeins gildan texta“ er að minnsta kosti jafn erfitt og tvígilda flokkunarverkefnið „er þessi staðhæfing gild?“.

Í öðru lagi færa þeir rök, studd könnun á **tíu áhrifamiklum viðmiðum**, fyrir því að hefðbundin nákvæmnisviðmið umbuni ágiskun frekar en fráhvarfi frá svari.

Í þriðja lagi leggja þeir til breytingu og sýna lítið dæmi: **opin matsviðmið** (*open rubrics*), þar sem stigareglan stendur inni í spurningunni sjálfri — til dæmis „rétt svar fær 1 stig, rangt −1, svo slepptu svari ef þú ert undir 50% viss“. Þá veit líkanið að heiðarleiki getur verið verðlaunaður.

Þeir prófa þetta á fjórum fremstu líkönum — Gemini 3 Pro frá Google, GPT-5 frá OpenAI, Grok 4 frá xAI og Claude Opus 4.5 frá Anthropic — með **4.326** staðreyndaspurningum úr SimpleQA. Höfundarnir leggja sjálfir áherslu á að þetta sé lýsandi dæmi, **ekki stýrður samanburður milli líkana**: sjálfgefnar stillingar, engin sértuning og engin normalisering fyrir kostnað.

Hvað fundu þeir?

- **Grunnþjálfun neyðir fram einhverja skekkju.** Tíðni sjálfsöruggra rangra staðhæfinga hefur formleg neðri mörk sem tengjast villutíðni besta flokkarans sem reynir að ákveða hvort staðhæfing sé gild. Fyrir staðreyndir án læranlegs mynsturs er lágmarkið tengt tíðni „einstæðinga“ — staðreynda sem koma aðeins einu sinni fyrir. Einhver skekkja getur því verið óhjákvæmileg jafnvel með hreinum gögnum.
- **Stigagjöf umbunar ágiskun í raun.** Með venjulegu rétt/rangt-skori er það hagstæðasta að sleppa aldrei svari, og könnun höfundanna sýnir að flest vinsæl viðmið skora „ég veit það ekki“ einfaldlega sem rangt. Á SimpleQA getur hrá nákvæmni til dæmis lítillega **haggað o4-mini** — sem svarar nánast öllu og hefur rangt fyrir sér í meira en þrjá fjórðu tilvika — fram yfir GPT-5-mini, sem gerir færri mistök vegna þess að það sleppir svari þegar það er óvisst. Reckless líkanið lítur betur út á stigatöflunni.
- **Opin matsviðmið snúa hvatanum við í litla dæminu.** Höfundarnir prófa einfalda aðferð: láta líkanið svara tvisvar og sleppa svari ef svörin eru ósammála. Undir venjulegri nákvæmni minnkar aðferðin villur **en einnig hráa nákvæmni**, þannig að viðmiðið refsar henni. Með opnu stigakerfi kemur sama aðferð betur út fyrir öll fjögur líkönin yfir bil af refsistigum. GPT-5-mini, sem hrá nákvæmni hafði refsað fyrir varfærni, fer líka fram úr o4-mini þegar stigareglurnar eru gerðar skýrar (n = 4.326 spurningar á líkan).

Hvað þýðir þetta líklega?

Að minnka rangar staðreyndir snýst líklega ekki fyrst og fremst um að búa til fleiri sérhæfð „hallucination-próf“. Það snýst um að breyta því hvernig **almennu viðmiðin** skora óvissu, þannig að „ég veit það ekki“ sé ekki sjálfkrafa refsað. Svo lengi sem stigataflan umbunar ágiskun mun aðferð sem minnkar rangsvör oft tapa hráum nákvæmnisstigum — og þar með verða verri kostur í keppninni.

Þess vegna kalla höfundarnir vandamálið **félags-tæknilegt**: hluti er betra mælikerfi og hluti er að fá áhrifamiklar stigatöflur til að nota það.

Hvað sýnir þetta *ekki*?

- Það sýnir **ekki** að opin matsviðmið leysi rangsvör í raunverulegri notkun. Tilraunin er lítil og viljandi **óstýrð**: fjögur líkön, sjálfgefnar stillingar, ein aðferð og eitt staðreyndapróf. Markmiðið er að sýna hvataáhrif, ekki að sanna almenna virkni eða raða líkönum.
- Það heldur **ekki** fram að stigagjöf sé **eina orsökin**. Villur í þjálfunargögnum, genuinely erfið verkefni og ókunnar fyrirspurnir eru aðrir sjálfstæðir uppsprettur.
- Það styður **ekki** þá vinsælu fullyrðingu að rangsvör séu óhjákvæmileg í öllum skilningi. Höfundarnir færa þvert á móti rök fyrir því að kerfi sem aðeins svaraði sannreynanlegum spurningum og segði annars „ég veit það ekki“ gæti forðast þessa tegund sjálfsöruggra rangra staðhæfinga.
- Það lætur **ekki** tölfræðilegu lágmörkin hverfa. Greinin útskýrir og afmarkar þau; þau varða sjálfsörugg staðreyndamistök, ekki alla hegðun líkans.
- Það sýnir **ekki** að opin matsviðmið séu **nægjanleg ein og sér**. Þau breyta því hvað próf verðlaunar; þau koma ekki í stað upplýsingaleitar, verkfæranotkunar eða betri kvörðunar líkansins.

Hversu sterk eru gögnin?

- Kjarninn er **stærðfræði** — formleg neðri mörk, ekki mælingar. Sem fræðileg röksemd stendur hann á eigin forsendum.
- Greinin notar **viljandi einfölduð líkön**. Höfundarnir nefna sjálfir „falska þrískiptingu“ þess að flokka öll svör sem rétt, röng eða „ég veit það ekki“, og nota hugsjónatilvik með handahófskenndum staðreyndum fyrir hreinustu mörkin.
- Könnunin á viðmiðum er **lítið, valið úrtak** — tíu áhrifamikil próf, ekki tæmandi úttekt á sviðinu.
- Dæmitilraunin er **raunveruleg en takmörkuð**: fjögur fremstu líkön, ein mótvægisaðferð, aðeins SimpleQA, sjálfgefnar stillingar og bein yfirlýsing um að þetta sé „not a controlled evaluation“.
- Sjónarhornið er líka rétt að nefna: þrír af fjórum höfundum starfa eða störfuðu hjá **OpenAI**, og greinin færi rök fyrir breytingu á því hvernig sviðið metur líkön. Það er rökstudd afstaða frá hagsmunaaðila, ekki hlutlaus utanaðkomandi úttekt. Hún verður metin, ekki sjálfkrafa hafnað. Til hróss beinist gagnrýnin líka að eigin líkönum OpenAI, o4-mini og GPT-5-mini.

Af hverju skiptir þetta máli?

Greinin endurrammar mjög ýkt vandamál. „Hallucination“ er oft kynnt annaðhvort sem dularfullur galli eða óhreyfanlegur veggur. Hér verður það hversdagslegra og að hluta sjálfskapað: tölfræðilegt lágmark sem við getum skilið, ofan á hvata sem við völdum sjálf. Víðari lærdómurinn er rólegri og gagnlegri: frekari framfarir í áreiðanleika geta ráðist jafn mikið af **því sem við mælum** og því sem við smíðum.

Í stuttu máli

Sjálfsörugg röng svör frá mállíkönum koma, samkvæmt greininni, frá tveimur aðskildum þáttum. Fyrri er tölfræðilegur: þegar staðreynd hefur ekkert mynstur sem hægt er að læra mun líkan sem lærir úr texta stundum hafa rangt fyrir sér, og hægt er að setja neðri mörk á þá skekkju, til dæmis út frá því hversu margar staðreyndir birtast aðeins einu sinni í þjálfun. Seinni þátturinn er hvati: næstum öll viðmið sem líkön eru raðað eftir gefa „ég veit það ekki“ sama skor og rangt svar, þannig að ágiskun getur alltaf borgað sig — jafnvel þannig að líkan sem er rangt í meirihluta svara slær varfærnara líkan sem sleppir svörum. Tillaga höfundanna er **opin matsviðmið**: setja stigaregluna inn í spurninguna. Í litlu dæmi með fjórum fremstu líkönum snýr það hvatanum þannig að aðferð sem minnkar rangsvör fær umbun í stað refsinga. Þetta er fræðileg grein með könnun og litlu, beinlínis óstýrðu dæmi, ritrýnd í *Nature*. Hugmyndin er lofandi en hefur ekki enn verið sýnd í stórum skala; meginboðið er að rangsvör eru hvorki dularfull né algerlega óhjákvæmileg.

Raunhæf yfirferð

Hvað sýnir greinin? Formleg neðri mörk á einhverja staðreyndaskekkju í grunnþjálfun, meðal annars tengd tíðni „einstæðinga“ fyrir mynsturlausar staðreyndir; könnun sem sýnir að flest áhrifamikil viðmið gefa „ég veit það ekki“ ekkert stig; og fjögurra-líkana dæmi þar sem opin stigaregla gerir aðferð sem minnkar rangsvör hagstæða þótt hrá nákvæmni hafi refsað henni.

Hvað er trúverðugt en ósannað? Að víðtæk upptaka opinna matsviðmiða í almennum viðmiðum myndi minnka rangsvör í raunverulegum kerfum. Tilraunin sem styður það er lítil og viljandi óstýrð.

Hvað sýnir greinin ekki? Að rangsvör séu óhjákvæmileg — hún færi rök fyrir hinu gagnstæða; að stigagjöf sé eina orsökin; að hægt sé að útrýma allri skekkju; eða að dæmitilraunin raði fjórum líkönunum almennt að gæðum.

Helstu takmarkanir: Viljandi einfalt rétt/rangt/„ég veit það ekki“-líkan sem höfundarnir kalla sjálfir „falska þrískiptingu“; lítið valið úrtak tíu viðmiða; óstýrt dæmi með fjórum líkönum, einni aðferð og einu prófi; og röksemd frá höfundahópi þar sem OpenAI er áberandi hagsmunaaðili.

Hversu mikið traust ætti almennur lesandi að leggja á niðurstöðuna? Mikið traust á að rangsvör séu hvorki dularfull né algerlega óhjákvæmileg og að algeng viðmið umbuni nú ágiskun. Miðlungs traust

á að tillagan um opin matsviðmið hjálpi í stærri kerfum — hún hefur raunverulega proof-of-concept niðurstöðu en ekki enn víðtæka staðfestingu.

Heimildir

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