



The CLEAN PAPER

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

Moodeelonni afaanii maaliif hallucinate godhu —
akkaataan qabxii itti kenninus maaliif akkasuma akka
turu godha

Sobni ofitti amanamuun himamu kan “hallucination” jennu dogoggora iccitii miti: tokko tokko bu’aa statistiksii leenjii ti; benchmarks baramaan tilmaama ofitti amanamu “hin beeku” amana-maa caalaa waan badhaasaniif hafan. Case study frontier models afur irratti, qabxii prompt keessa ibsuun (“open rubrics”) kaka’umsa sana akka garagalchu agarsiisa.

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

Moodeelonni maaliif tilmaamu, nutis maaliif akkas akka godhan barsiifne

Large language model tokko guyyaa dhaloota nama hin beeknee gaafadhu; “March 7” jedhee amanamummaa nama kaardii irraa dubbisuu fakkaatuun deebisuu danda’a — yeroo sadii walitti aansuun, guyyoota adda addaa sadiin, dogoggoruus danda’a. Barreessitoonni fakkeenya akkanaa sirriitti kennu: moodeelota dura deemoo gaaffii dhugaa salphaa — guyyaa dhaloota nama tokkoo yookaan acronym hin beekamne tokko maal bakka bu’a — gaafatan; tokkoon tokkoon isaanii deebii adda addaa ofitti amanamuun uuman, tokko illee sirrii miti. Jechi industiriin kanaaf fayyadamu *hallucination* dha; kunis akka rakkoo waan argan hubachuu keessaa fakkeessa. Tarkaanfiin barruu kanaa jalqabaa iccitii sana keessaa baasuu dha.

Akka moodeelli ijaaramu irraa jalqabi. Sadarkaa leenjii isaa jalqabaa fi hunda caalaa guddaa keessatti, barruu hedduu dubbisuun afaan sirriitti yaa’u maal akka fakkaatu barataa. Amma dhugaa pattern tokko illee duuba hin qabne fudhadhu — guyyaa dhaloota nama tokkoo. Guyyaan sun barruu leenjii keessatti yeroo tokko yoo mul’ate, yookaan matumaa yoo hin mul’anne, wanti barataan pattern qabachuu danda’u hin jiru: deebiin sun, ilaalcha moodeelichaan, akkasumaan filatame. Barreessitoonni yaada durii (kan Alan Turing, rakkoo biraa irraa) fudhachuun kana sirriitti ibsu: guyyaa dhalootaa shan keessaa tokko daataa keessatti yeroo tokko qofa yoo mul’ate, moodeelli yoo xiqqaate shan keessaa tokko irratti dogoggora jedhamee eegamuu qaba — sababa inni cabee miti, sababa wanti baratamu achi hin turreef. (Yaada wal fakkaatuun, moodeelonni magaalaa guddoo biyya tokkoo irratti yeroo muraasa qofa dogoggoru: isaan yeroo hundaa mul’atu.) Himata dhugaa fi soba dhugaa fakkaatu adda baasuun mataan isaa rakkoo ulfaataa akka ta’e, himata dhugaa *qofa* uumuun ammoo yoo xiqqaate sanuma akka ulfaatu of eeggannoodhaan falmu. Daangaan dogoggoraa tokko sirnicha keessa duraan jira.

Yaada jalaa jiru: waan amma iyyuu hin argine akkamitti lakkoofa?

Kun yaada dhugumaan qaroo — moodeelota afaanii caalaa dulloome, sirriitti beekuun gatii qabu — irratti hundaa’a.

Boorsaa kubbaa halluu qabu irraa jalqabi. Halluu meeqa akka qabu hin beektu. Tokko tokkoon 100 keessaa baaftte lakkoofa: diimaa 40, cuquliisa 25, magariisa 15, keelloo 5, diimaa dhiilgee 3, burtukaana 2 — achiis **halluu adda addaa kudhan, tokkoon tokkoon isaanii yeroo tokko qofa mul’atan.**

Amma gaaffii Turing rakkoo guutummaatti adda ta’e irratti dhugumaan mudate: *kubbaawwan itti aanan keessaa tokko halluu matumaa hin argine ta’uu carraan isaa maali?* Waan matumaa hin baafne lakkaa’uu hin dandeessu — garuu halluwwan yeroo tokko qofa argite, “singletons,” **lakkaa’uu dandeessa.** Mala **Good-Turing estimation** jedhamu keessatti, qoodni draws kee singletons ta’an probability halluwwan hin argine keessa dhokate tilmaama. Draws dhibba keessaa kudhan halluu yeroo tokko qofa turan; kanaaf kubbaan itti aanu halluu guutummaatti haaraa ta’uu carraan isaa tilmaamaan **10 / 100 = 10%.**

Halluwwan yeroo tokko argaman sun *dogoggora* miti. Isaan wallaalummaa mataa keetii safaru: halluwwan hedduun yeroo tokko mul’achuun sample sun wanti ati amma iyyuu hin baafne hedduun addunyaa keessa akka jiru sitti himuu dha.

Amma halluu bakka guyyaa dhalootaa, boorsaa bakka barruu leenjii moodeelichaa buusi. Guyyoota dhalootaa inni arge keessaa **shan keessaa tokko yeroo tokko qofa mul'ata** haa jennu. Mala wal fakkaatu: probability keessaa tilmaamaan harka shan keessaa tokko guyyaa dhalootaa moodeelichi akka waan hin arginetti ilaalamu keessa jira — guyyaan dhalootaa pattern ittiin duubatti deebi'amu hin qabu (guyyaa dhaloota nama tokkoo *shallaguu* hin dandeessu). Kanaaf guyyaan yeroo tokko argame yookaan matumaa hin argamne saantima moodeelichi ulfaatina itti kennuu hin dandeenye dha; isaan keessaa tilmaamaan **shan keessaa tokko** irratti dogoggora. Qaroomni kam illee kana hin suphu: wanti baratamu hin turre. Falmii guutuun bifa xiqqaa kana: rate singleton qoodni addunyaa *daataa kana irraa* barachuun hin danda'anne hammam akka ta'e safara; kunis dogoggora jalatti **daangaa gadi-aanaa** ta'a. Moodeelli magaalaa guddoo tokko maaliif yeroo muraasa qofa dogoggorus ibsa — *Paris* yeroo hundaa mul'ata, rate singleton isaa zeerootti dhihaata; kanaaf wanti baratamu baay'ee jira.

Kun hallucinations eessaa akka dhufan ibsa. Maaliif akka isaan *hafan* — leenjii booda moodeelota nama gargaaran fi amanamoo gochuuf yaalame hunda boodas, maaliif shakkii isaanii himuu mannaa akka waan beekaniitti dubbatan — hin ibsu. Asitti wal fakkeenya barruu kanaa nama qaanessuutti dhihaata. Barataa qorannoo keessatti deebii hin beekne yaadi. Wanti duwwaan zeero argatee, tilmaamni tokko argachuu danda'a yoo ta'e, tarkaanfiin qabxii guddisu tilmaamuu dha — ofitti amanamummaadhaan, sirriitti, “hin mirkaneeffadhu” osoo hin jedhin. Barattoonni kana baratu. Moodeelonni ni baratu — sababni isaas karaa wal fakkaatuun qabxii isaanii kennina. Barreessitoonni benchmarks dameen kun irratti dorgomu, leaderboards moodeelonni ol ba'uuf qindaa'an, keessa darban; jechuun ni danda'ama isaan hundi “hin beeku” fi deebii dogoggoraa qabxii wal fakkaatu, zeero, akka kennan argan. Seera sana jalatti, moodeelli yeroo hundaa tilmaamu moodeela karaa biraatiin guutummaatti wal fakkaatu garuu shakkii isaa amanamummaadhaan ibsu ni caala. Hiika kallattiitti dhihaatuun, qabxii kennuun akkas akka godhan isaan barsiifna.

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."

Benchmarks tilmaamuun sirrii akka fakkaatu gochuu danda'u. Qabxii benchmarks irra caalaan fayyadamu jalatti (bitaa), deebiin dogoggoraa fi "hin beeku" amanamaan lamaan isaanii zeero argatu — kanaaf tilmaamni gargaaruu qofa danda'a. Seera gaaffii keessa, dogoggoraaf adabbii wajjin, ibsi (mirga); yeroo hin mirkanoofne deebii dhiisuun filannoo gaarii ta'uu danda'a. Wanta qorannoon badhaasu jijjiira; ofuma isaatiin hallucination hin furu.

Original diagram — The Clean Paper CC BY 4.0

Kutaan kun akka eegamu gatii qaba, sababiin isaas mata duree baramaa faallaa deema. Hallucination yeroo baay'ee akka daangaa technology irraa adda hin baane, iccitii fakkaatutti dhiyaata. Barruun kun lamaan isaanii morma. Daangaan pretraining iccitii miti — dogoggora statistiksii idilee, gosa machine learning waggoota kudhan hedduuf hubate dha. Hafuun isaas dirqama miti — gartokkoon isaa kaka'umsa nuti ijaarree jijjiiruu dandeenyu dha. Sirni yeroo hin mirkanoofne deebii kennuu didu qofti matumaa hallucinate hin godhu; moodeelonni hojii irra jiran akkas hin amalle sababiin scoreboard keenya deebii dhiisu adabuufi.

Barreessitoonni maal godhan?

Barruun kun kutaa sadii qaba. Tokkoffaa, pretraining keessatti hallucination tokko tokko statistiksii dhaan dirqama ta'uu isaa falmii herregaa, "barruu sirrii qofa uumi" jechuun yoo xiqqaate rakkoo binary classification "himannaan kun sirrii dha?" jedhuu wajjin wal-qixa ulfaataa akka ta'e agarsiisuun. Lammaffaa, falmii — qorannoo benchmarks dhiibbaa qaban kudhaniin deeggarama — metrics accuracy baramaan tilmaamuun deebii dhiisuu caalaa akka badhaasan. Sadaffaa, furmaata yaadame fi case study isa qoru: qorannoo **open-rubric**, bakka sirni qabxii gaaffii mataa isaa keessatti ibsamu (fakkeenyaaf, "deebiin sirrii 1 argata, kan dogoggoraa -1, kanaaf 50% gadi yoo mirkaneeffatte dee-

bii dhiisi”), akka moodeelli amanamummaan yoom badhaafamu beekuu danda’utti. Kana frontier models afur — Gemini 3 Pro kan Google, GPT-5 kan OpenAI, Grok 4 kan xAI fi Claude Opus 4.5 kan Anthropic — irratti gaaffilee dhugaa 4,326 kan SimpleQA fayyadamuun yaalan. Case study sun fakkeenya akka ta’e, “qorannoo moodeelota gidduutti to’atame miti” (default settings, tuning hin qabu, cost normalisation hin qabu), ifatti dubbatu.

Maal argan?

- **Pretraining dogoggora tokko tokko dirqisiisa.** Rate moodeelli soba ofitti amanamuun baasu, error rate classifier “himannaan kun sirrii dha?” isa irraa ijaarame keessaa hunda caalaa gaarii ta’ee (tilmaamaan dachaa lama) gadi hin bu’u. Dhugaawwan pattern baratamu hin qabneef, daangaan kun yoo xiqqaate *rate singleton* — qooda dhugaawwan leenjii keessatti yeroo tokko qofa mul’atan — dha. Daataan guutummaatti qulqulluu ta’us hallucination tokko tokko hambisuun hin danda’amu.
- **Qabxii kennuun tilmaamu badhaasa — ifatti.** Sirna sirrii/dogoggoraa baramaa jalatti, matumaa deebii dhiisuu dhabuun strategy gaarii dha; qorannoon barreessitootaa benchmarks beekamoo irra caalaan “hin beeku” akka dogoggoraatti lakkaa’an argata. Fakkeenyi isaanii keessaa ifaan: qorannoo SimpleQA irratti, raw accuracy xiqqoo *o4-mini kan OpenAI filata* — inni waan hunda jechuun ni danda’ama deebisee yeroo harka sadii keessaa afur caalaa dogoggora — GPT-5-mini, shakkii yeroo qabu deebii dhiisuun dogoggora baay’ee xiqqeessu, caalaa. Moodeelli balaa guddaa fudhatu scoreboard irratti gaarii fakkaata.
- **Open rubrics kaka’umsa garagalchu (case study isaanii keessatti).** Mala hallucination salphaa xiqqeessu tokko qoru (moodeelli yeroo lama deebisee, deebiin lamaan yoo wal hin simne akka dhiisu). Accuracy baramaa jalatti, malli sun dogoggora *xiqqeessa garuu accuracy illee xiqqeessa* — kanaaf metric sun itti fayyadamuu dadhabsiisa. Open rubrics jalatti, malli wal fakkaatu models afur hundaaf adabbii adda addaa keessatti caala; GPT-5-mini — raw accuracy deebii shakkii yeroo qabu dhiisuu isaaf adabe — qabxiin ifatti yeroo ibsamu o4-mini caala (n = 4,326 gaaffii moodeela tokkoof).

Kun maal jechuu ta’uu danda’a?

Hallucination xiqqeessuun irra caalaa qorannoo hallucination addaa dabalataa uumuun miti. Benchmarks baramaan uncertainty akkamitti akka qabxii kennan jijjiiruun, “hin beeku” jedhanii amanamummaadhaan himuun akka hin adabamne gochuu dha. Hanga scoreboard jijjiiramutti, hallucination xiqqeessuun qabxii accuracy moodeelotaa keessaa hir’isa, kanaaf ni dadhabsiifama — barreessitoonni rakkoo kana “socio-technical” jechuun isaanii kana: gartokkoon metric gaarii, gartokkoon leaderboards dhiibbaa qaban akka fudhatan gochuu.

Kun maal *hin mirkaneessu*?

- Open rubrics ayyaana hojii keessatti hallucination akka suphu **hin agarsiisu**. Qorannoon deeggaru case study xiqqaa, itti yaadamee **hin to’atamne** — moodeelota afur default settings irratti, mala tokko, factual-QA test tokko — kan *kaka’umsa garagaluu* agarsiisuuf ta’e; moodeelota sadarkaa kaa’uuf yookaan bu’a-qabeessummaa waliigalaa mirkaneessuuf miti.
- Qabxii kennuun *sababa qofa* akka ta’e **hin jedhu**. Dogoggora daataa leenjii, rakkoowwan dhugu-maan ulfaatoo fi prompts haaraan maddoota adda addaa ta’anii hafu.
- Hallucinations dirqama jedhu baramaa **hin deeggaru**. Barreessitoonni faallaa isaa falmu: sirni gaaffilee mirkanaa’uu danda’an qofa deebisee, kan biroo “hin beeku” jedhu, matumaa hallucinate hin godhu.
- Daangaa pretraining **hin balleessu** — ibsa, daangessas; daangaan sun dogoggora dhugaa ofitti amanamuun himamu irratti, amala moodeela hunda irratti miti.
- Open rubrics ofuma isaatiin *gahaa* ta’uu **hin agarsiisu**. Wanta qorannoon badhaasu jijjiiru; retrieval, tool use yookaan moodeelota caalaatti calibrated ta’an hin bakka bu’u.

Ragaan hammam cimaa dha?

- Bu’uuri isaa **herrega** dha — formal lower bounds, safartuu miti. Falmii theory ta’uun daangaa yaadota isaa keessatti cimaa dha.
- Rakkoo **itti yaadamee salphifame** irratti hundaa’a; barreessitoonni deebii hunda sirrii, dogoggoraa yookaan “hin beeku” jedhamutti qooduun “false trichotomy” ta’uu fi setting “arbitrary facts” daangaa qulqulluuf fayyadaman ni ibsu.
- Qorannoon benchmark **sample xiqqaa filatame** dha — evaluations dhiibbaa qaban kudhan, audit guutuu miti.
- Case study **dhugaa garuu daangeffame** dha: frontier models afur, mitigation tokko, SimpleQA qofa, default settings; ifatti “qorannoo to’atame miti.” Falmii kaka’umsaaf proof of concept dha, bu’aa benchmark miti.
- Ilaalchi eessaa akka dhufe maqaa dhahuun gatii qaba: barreessitoota afur keessaa sadii **OpenAI** keessa hojjetu yookaan hojjetaniiru; barruunis dameen kun akkaataa moodeelota itti madaalu akka jijjiiru falma. Ilaalcha namoota dhimma keessatti qooda qaban irraa falmii gaariidhaan dhiyaate dha, madaallii ala irraa walaba ta’e miti — madaaluuf, tuffachuuf miti. (Gaariin isaa, qeeqa sana moodeelota isaanii o4-mini fi GPT-5-mini irrattiis akkuma kan biroo irratti qajeelcha.)

Maaliif barbaachisaa dha?

Rakkoo garmalee beeksifame bifa haaraan kaa’a. “Hallucination” akka dogoggora sodaachisaa yookaan dallaa hin sochooneetti dhiyaata; barruun kun waan idilee fi gartokkoon nuti ofumaan uumnetti deebisa — daangaa statistiksii hubachuu dandeenyu, kaka’umsa nuti filanne irra taa’e. Barumsi bal’aan tasgabbaa’aa fi caalaatti fayyadu: reliability irratti tarkaanfiin itti aanu *waan safarru*

irratti akkuma waan ijaarru irratti hundaa’uu danda’a.

Cuunfaa ifaa

Deebiin sobaa moodeelli afaanii ofitti amanamummaadhaan kennu bakka lama irraa dhufa. Tokkoffaan statistiksii dha: dhugaan pattern baratamu yoo hin qabne, moodeelli afaan fakkeessuuf leenjifame yeroo tokko tokko dogoggora; daangaan sunis tilmaamamuu danda’a (fakkeenyaaf, dhugaawwan leenjii keessatti yeroo tokko qofa mul’atan meeqa akka ta’an irraa). Lammaffaan kaka’umsa: benchmarks moodeelota sadarkaa kaa’an jechuun ni danda’ama hundi “hin beeku” fi deebii dogoggoraa qabxii wal fakkaatu kennu; kanaaf tilmaamuun yeroo hundaa caala — moodeelli yeroo harka sadii keessaa afur dogoggoru kan amanamummaan deebii dhiisu caaluu danda’a. Yaadni barreessitootaa qorannoo hallucination biraa miti; “open rubrics” dha: qabxii gaaffii keessatti ibsi. Case study frontier models afur irratti, kun kaka’umsa garagalchee mala hallucination xiqqeessu adabuu mannaa badhaasa. Barruu theory-fi-qorannoo, qorannoo xiqqaa ifatti hin to’atamne, peer-reviewed *Nature* keessatti ba’e dha; furmaatni abdachiisaa dha, garuu bal’inaan hojjechuun isaa amma iyyuu hin mirkanoofne; hallucination iccitii yookaan guutummaatti dirqama akka hin taane falma.

Sakatta’a soba-maleessa

Barruun maal agarsiisa: Pretraining keessatti hallucination tokko tokko dirqisiisuuf lower bound herregaa (dhugaawwan pattern hin qabneef yoo xiqqaate “rate singleton”); qorannoo benchmarks dura deemoo irra caalaan “hin beeku”f qabxii akka hin kennine argate; fi case study moodeelota afur keessatti qabxii prompt keessa ibsuun (“open rubrics”) mala hallucination xiqqeessu caalchise, bakka accuracy salphaan isa adabe.

Wanti ta’uu danda’u garuu hin mirkanoofne: Open rubrics benchmarks baramaa keessa yoo galan moodeelota hojii irra jiran keessatti hallucination hiika qabuun xiqqeessuu. Qorannoon deeggaru xiqqaa fi ifatti hin to’atamne.

Wanti inni hin agarsiisne: Hallucinations dirqama ta’uu (faallaa isaa falma); qabxii kennuun sababa qofa ta’uu; hallucination guutummaatti dhabamsiisuu danda’amu; case study sun models afur wal bira qabuun sadarkaa kaa’uu.

Daangaawwan ijoo: Moodeela sirrii/dogoggoraa/“hin beeku” jedhamee salphifame (barreessitoonni “false trichotomy” jedhu); benchmark sample xiqqaa filatame (evaluations kudhan); case study hin to’atamne (models afur, mitigation tokko, test tokko, default settings); fi falmii OpenAI’n durfamu akka dameen moodeelota madaalu irratti.

Dubbisaan waliigalaa hammam amanuu qaba? Hallucination iccitii yookaan guutummaatti dirqama akka hin taane, benchmarks baramaan ammoo tilmaamu akka badhaasan irratti amantaa guddaa. Furmaatni yaadame gargaara irratti amantaa giddu-galeessa — proof of concept dhugaa

amma jira, garuu bal'inaan fi scale irratti hojjechuun isaa hin agarsiifamne.

Maddoota

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>