The Clean Paper

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

Чаму моўныя мадэлі галюцынуюць — і чаму наша сістэма ацэнкі дапамагае гэтаму працягвацца

Упэўненыя ілжывыя сцвярджэнні, якія мы называем «галюцынацыямі», не з'яўляюцца загадкавым збоем: частка з іх — статыстычны пабочны эфект навучання, а захоўваюцца яны яшчэ і таму, што галоўныя бенчмаркі ўзнагароджваюць упэўненую здагадку мацней за сумленнае «я не ведаю». Дэманстрацыйнае даследаванне на чатырох frontier-мадэлях паказвае, што яўнае фармуляванне правілаў ацэнкі ў самім запыце — «open rubrics» — пераварочвае гэты стымул.

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

Чаму мадэлі ўгадваюць — і чаму мы самі іх гэтаму навучылі

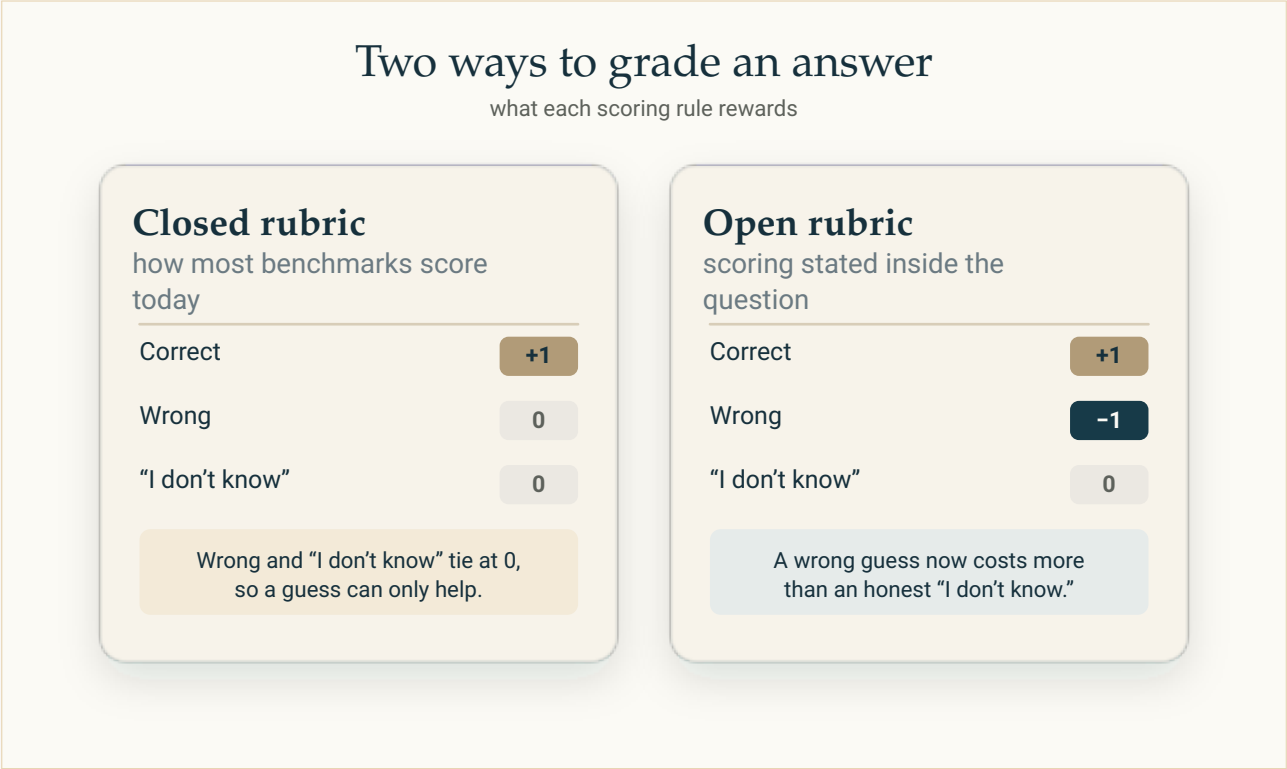
Спытайце вялікую моўную мадэль пра дзень нараджэння незнаёмага чалавека — і яна можа адказаць «7 сакавіка» са спакойнай упэўненасцю таго, хто чытае дату з карткі, а потым тры разы запар памыліцца, кожны раз называючы іншую дату. Аўтары прыводзяць менавіта такія прыклады: вядучым мадэлям задаюць простае фактычнае пытанне — дзень нараджэння чалавека або расшыфроўку малавядомай абрэвіятуры — і кожная ўпэўнена прыдумляе свой адказ, прычым ніводны не правільны. У галіне гэта называюць *галюцынацыяй*, нібы гаворка ідзе пра збой успрымання. Першы крок артыкула — прыбраць з гэтай з'явы містыку.

Пачнём з таго, як будуюць мадэль. На першым і найбуйнейшым этапе навучання яна, па сутнасці, вучыцца таму, як выглядае натуральная мова, чытаючы велізарны масіў тэксту. Цяпер возьмем факт без заканамернасці — канкрэтны дзень нараджэння пэўнага чалавека. Калі гэтая дата сустрэлася ў навучальным тэксце адзін раз або не сустрэлася зусім, мадэлі, якая вучыцца на патэрнах, няма за што зачапіцца: з яе пункту гледжання адказ адвольны. Аўтары робяць гэтую ідэю дакладнай, выкарыстоўваючы стары падыход Алана Цьюрынга з іншай задачы: калі кожны пяты дзень нараджэння сустракаецца ў даных толькі адзін раз, варта чакаць, што мадэль будзе памыляцца прынамсі прыкладна ў адным выпадку з пяці — не таму, што яна «зламаная», а таму, што ў даных не было заканамернасці, якую можна вывучыць. (Паводле той жа логікі мадэлі амаль ніколі не памыляюцца са сталіцамі краін: гэтыя факты паўтараюцца пастаянна.) Аўтары акуратна даказваюць, што адрозніць праўдзівае сцвярджэнне ад праўдападобнага ілжывага — само па сабе складаная задача, а генерыраваць *толькі* праўдзівыя сцвярджэнні прынамсі не прасцей. Значыць, пэўная ніжняя мяжа памылак закладзена статыстычна.

> **Ідэя пад гэтым вынікам: як палічыць тое, чаго вы яшчэ не бачылі**
>
> Тут выкарыстоўваецца сапраўды прыгожая ідэя — старэйшая за моўныя мадэлі і вартая асобнага знаёмства.
> **Пачнём з мяшка каляровых шарыкаў.** Вы не ведаеце, колькі ў ім колераў. Вы выцягваеце 100 шарыкаў па адным і лічыце: чырвоных 40, сініх 25, зялёных 15, жоўтых 5, фіялетавых 3, аранжавых 2 — а затым **дзесяць розных колераў, кожны з якіх сустрэўся роўна адзін раз**.
> Цяпер пытанне, з якім Цьюрынг насамрэч сутыкнуўся ў зусім іншай задачы: *якая імавернасць, што наступны шарык будзе колеру, якога мы ўвогуле яшчэ не бачылі?* Палічыць тое, чаго вы ніколі не выцягвалі, немагчыма — але **можна** палічыць колеры, якія сустрэліся роўна адзін раз, гэта значыць «singletons». Прыём, вядомы як **ацэньванне Гуда–Цьюрынга (Good–Turing estimation)**, палягае ў тым, што доля аднаразовых назіранняў ацэньвае імавернасную масу, якая ўсё яшчэ хаваецца ў невядомых катэгорыях. Дзесяць са ста выцягнутых шарыкаў належалі колерам, якія сустрэліся толькі раз, значыць, імавернасць таго, што наступны колер будзе зусім новым, складае прыкладна **10 / 100 = 10%.**

> Гэтыя аднаразовыя колеры — не *памылкі*. Яны з’яўляюцца вымярэннем вашага ўласнага няведання: вялікая колькасць колераў, якія з’явіліся адзін раз, — спосаб, якім выбарка паведамляе, што ў свеце ёсць яшчэ шмат таго, чаго вы проста не выцягнулі.
> **Цяпер заменім колеры на дні нараджэння,** а мяшок — на навучальны тэкст мадэлі. Дапусцім, сярод убачаных дзён нараджэння **кожны пяты сустракаецца роўна адзін раз.** Той самы прыём: прыкладна пятая частка імавернаснай масы прыпадае на дні нараджэння, якіх мадэль фактычна ніколі не бачыла, а ў дня нараджэння няма заканамернасці, на якую можна абаперціся — яго нельга *вылічыць*. Таму дата, убачаная адзін раз або ніколі, — манета, вагу якой мадэль не ведае, і яна будзе памыляцца прыкладна ў **адным выпадку з пяці**. Ніякая «разумнасць» не выправіць адсутнасці інфармацыі: вучыцца не было на чым.
> У мініяцюры гэта ўвесь аргумент: доля singletons вымярае, колькі свету *немагчыма вывучыць з гэтых даных*, і гэта становіцца **ніжняй мяжой** памылак. Менавіта таму мадэль амаль ніколі не промахваецца са сталіцай: *Paris* сустракаецца пастаянна, яго singleton rate блізкі да нуля, значыць, інфармацыі для навучання шмат.

Гэта тлумачыць, адкуль бяруцца галюцынацыі. Але не тлумачыць, чаму яны *выжываюць* — чаму пасля ўсіх наступных этапаў навучання, прызначаных зрабіць мадэль карыснай і сумленнай, яна ўсё роўна блефуе замест прызнання няўпэўненасці. Тут аналогія аўтараў амаль няёмка дакладная. Уявіце студэнта на экзамене, які не ведае адказу. Калі пусты адказ дае нуль, а здагадка можа даць адзін бал, стратэгія, якая максімізуе ацэнку, — угадваць: упэўнена, канкрэтна, ніколі не пісаць «я не ведаю». Студэнты гэтаму вучацца. Як высвятляецца, мадэлі таксама — бо мы ацэньваем іх гэтак жа. Аўтары прагледзелі бенчмаркі, на якіх рэальна спаборнічае галіна, рэйтынгі, пад якія мадэлі аптымізуюць, і выявілі: амаль усе даюць «я не ведаю» роўна столькі ж балаў, колькі няправільнаму адказу, — нуль. Пры такім правіле мадэль, якая заўсёды ўгадвае, абгоніць ідэнтычную мадэль, якая сумленна пазначае няўпэўненасць. У даволі літаральным сэнсе мы самі сістэмай ацэнкі падштурхоўваем іх да гэтага.



Бенчмаркі могуць рабіць угадванне рацыянальнай стратэгіяй. Пры схеме ацэнкі, якую выкарыстоўвае большасць тэстаў (злева), няправільны адказ і сумленнае «я не ведаю» абодва даюць нуль — значыць, здагадка можа толькі дапамагчы. Калі правілы яўна запісаць у самім пытанні і штрафаваць за памылку (справа), адмова ад адказу пры няўпэўненасці можа стаць лепшым выбарам. Гэта мяняе тое, што ўзнагароджвае тэст; само па сабе яно не вырашае праблему галюцынацый.

Original diagram — The Clean Paper CC BY 4.0

Менавіта гэтую частку варта запомніць, бо яна супярэчыць звыкламу загалоўку. Галюцынацыі часта падаюць як непазбежную, амаль містычную мяжу тэхналогіі. Артыкул пярэчыць абедзвюм часткам. Ніжняя мяжа на этапе pretraining не загадка — гэта звычайная статыстычная памылка, якую машыннае навучанне разумее дзесяцігоддзямі. А яе захаванне не непазбежнае: часткова гэта стымул, які мы самі стварылі і можам змяніць. Сістэма, якая проста адмаўляецца адказваць, калі не ўпэўнена, увогуле не галюцынавала б; разгорнутыя мадэлі не паводзяць сябе так таму, што нашы табліцы вынікаў караюць адмову.

Што зрабілі аўтары

Артыкул складаецца з трох частак. Першая — матэматычны аргумент, што пэўная доля галюцынацый статыстычна непазбежная падчас папярэдняга навучання: аўтары паказваюць, што задача «генерыраваць толькі карэктны тэкст» прынамсі такая ж складаная, як бінарная класіфікацыя «гэта сцвярджэнне карэктнае ці не?». Другая — аргумент, падмацаваны аглядам дзесяці ўплывовых бенчмаркаў, што звычайныя метрыкі дакладнасці ўзнагароджваюць угадванне мацней за ўстрыманне ад адказу.

Трэцяя — прапанаванае выпраўленне і дэманстрацыйнае даследаванне: **open-rubric** ацэньванне, дзе правіла падліку балаў проста запісана ў пытанні (напрыклад, «правільны адказ дае 1 бал, няправільны −1, таму ўстрымайцеся, калі ваша ўпэўненасць ніжэй за 50%»), каб мадэль ведала, калі сумленнасць узнагароджваецца. Гэта правяраюць на чатырох frontier-мадэлях — Google's Gemini 3 Pro, OpenAI's GPT-5, xAI's Grok 4 і Anthropic's Claude Opus 4.5 — выкарыстоўваючы 4 326 фактычных пытанняў SimpleQA. Аўтары наўпрост пішуць, што дэманстрацыйнае даследаванне толькі ілюстрацыйны і «не з'яўляецца кантраляваным параўнаннем мадэляў» — стандартныя налады, без т'юнінгу і нармалізацыі кошту.

Што яны выявілі

- **Pretraining прымушае мець пэўную долю памылак.** Частата, з якой мадэль выдае ўпэўненыя ілжывыя сцвярджэнні, мае ніжнюю мяжу, прыблізна звязаную з падвоенай частатой памылкі найлепшага класіфікатара «ці карэктнае гэта сцвярджэнне?», пабудаванага з яе. Для фактаў без заканамернасці, якую можна вывучыць, гэтая мяжа прынамсі роўная *singleton rate* — долі фактаў, якія сустракаюцца ў навучанні роўна адзін раз. Некаторыя галюцынацыі непазбежныя нават пры цалкам чыстых даных.
- **Ацэньванне канкрэтна ўзнагароджвае ўгадванне.** Пры звычайнай схеме «правільна/няправ аптымальная стратэгія — ніколі не ўстрымлівацца ад адказу, а агляд аўтараў паказвае, што пераважная большасць папулярных бенчмаркаў залічвае «я не ведаю» проста як памылку. Яркі прыклад з іх уласных мадэляў: на SimpleQA сырая accuracy крыху *аддае перавагу* OpenAI o4-mini, якая адказвае амаль на ўсё і памыляецца больш чым у трох чвэрцях выпадкаў, перад GPT-5-mini, якая робіць значна менш памылак, бо ўстрымліваецца, калі не ўпэўнена. Больш безразважная мадэль выглядае лепш у табліцы.
- **Яўныя крытэрыі ацэнкі (*open rubrics*) мяняюць стымул — гэта паказвае дэманстрацыйнае даследаванне.** Аўтары тэстуюць просты спосаб памяншэння галюцынацый: папрасіць мадэль адказаць двойчы і ўстрымацца, калі два адказы не супадаюць. Пры стандартнай метрыцы дакладнасці метад памяншае памылкі, *але адначасова зніжае паказчык дакладнасці*, таму метрыка фактычна карае яго выкарыстанне. Пры яўных крытэрыях ацэнкі той самы метад аказваецца лепшым для ўсіх чатырох мадэляў у шырокім дыяпазоне штрафаў; а GPT-5-mini, якую сырая accuracy карала за адмову ад адказу пры няўпэўненасці, абганяе o4-mini, калі схема балаў сфармулявана адкрыта (n = 4 326 пытанняў на мадэль).

Што гэта, верагодна, азначае

Зніжэнне галюцынацый у асноўным не патрабуе прыдумляць яшчэ больш спецыяльных тэстаў на галюцынацыі. Трэба змяніць тое, як галоўныя бенчмаркі ацэньваюць

няўпэўненасць, каб прызнанне «я не ведаю» перастала карацца. Пакуль табліца вынікаў не зменіцца, памяншэнне галюцынацый па-ранейшаму будзе каштаваць мадэлям балаў accuracy і таму заставацца нявыгадным — менавіта таму аўтары называюць праблему «сацыятэхнічнай»: патрэбныя і лепшая метрыка, і гатоўнасць уплывовых рэйтынгаў яе прыняць.

Чаго гэтая праца *не* даказвае

- Яна **не** паказвае, што яўныя крытэрыі ацэнкі выпраўляюць галюцынацыі ў рэальным выкарыстанні. Падтрымліваючы эксперымент — невялікі і наўмысна **некантраляванае** дэманстрацыйнае даследаванне: чатыры мадэлі са стандартнымі наладамі, адзін выбраны спосаб памяншэння галюцынацый, адзін тэст фактычных пытанняў. Яго мэта — паказаць *пераварот стымулу*, а не ранжыраваць мадэлі або даказаць агульную эфектыўнасць.
- Яна **не** сцвярджае, што ацэньванне — *адзіная* прычына. Памылкі ў навучальных даных, сапраўды складаныя задачы і незнаёмыя запыты застаюцца асобнымі крыніцамі.
- Яна **не** падтрымлівае папулярнае сцвярджэнне, што галюцынацыі непазбежныя. Аўтары сцвярджаюць супрацьлеглае: сістэма, якая адказвае толькі на правяральныя пытанні і інакш кажа «я не ведаю», ніколі не галюцынавала б.
- Яна **не** прыбірае ніжнюю мяжу pretraining — яна яе тлумачыць і абмяжоўвае, а сама мяжа датычыцца ўпэўненых фактычных памылак, не ўсіх паводзін мадэлі.
- Яна **не** паказвае, што саміх па сабе яўных крытэрыяў ацэнкі *дастаткова*. Яны мяняюць тое, што ўзнагароджвае ацэньванне; яны не замяняюць пошук інфармацыі, выкарыстанне інструментаў або лепш адкалібраваныя мадэлі.

Наколькі пераканаўчыя доказы?

- Ядро працы — **матэматыка**: фармальныя ніжнія межы, а не вымярэнні. Як тэарэтычны аргумент яна паслядоўная ў межах уласных дапушчэнняў.
- Яна абапіраецца на **наўмысна спрошчаныя мадэлі** праблемы. Самі аўтары адзначаюць «ілжывую трыхатамію» — падзел кожнага адказу толькі на правільны, няправільны або «я не ведаю» — і ідэалізаванае асяроддзе «адвольных фактаў», выкарыстанае для самай чыстай мяжы.
- Агляд бенчмаркаў — **невялікая, уручную адабраная выбарка**: дзесяць уплывовых ацэнак, а не поўны аўдыт.
- Дэманстрацыйнае даследаванне **рэальны, але абмежаваны**: чатыры frontier-мадэлі, адзін метад змяншэння галюцынацый, толькі SimpleQA, стандартныя налады, і наўпрост сказана «не кантраляванае параўнанне». Гэта proof of concept для аргумента пра стымулы, а не новы рэйтынг мадэляў.
- Варта назваць і пазіцыю аўтараў: трое з чатырох працуюць або працавалі ў **OpenAI**,

а артыкул аргументуе, як галіне варта змяніць ацэньванне мадэляў. Гэта добра аргументаваная пазіцыя зацікаўленага боку, а не нейтральны вонкавы агляд — яе варта ўзважваць, а не адкідаць. (Да гонару артыкула, ён скіроўвае крытыку на ўласныя мадэлі o4-mini і GPT-5-mini гэтак жа ахвотна, як і на іншыя.)

Чаму гэта важна

Праца пераасэнсоўвае надзвычай раскручаную праблему. «Галюцынацыю» часта прадаюць або як жудасны дэфект, або як нерухомую сцяну; гэты артыкул робіць яе больш звычайнай і часткова самаставранай — статыстычная ніжняя мяжа, якую мы можам зразумець, плюс стымул, які мы самі выбралі. Больш шырокі ўрок цішэйшы і карыснейшы: далейшы прагрэс у надзейнасці можа залежаць не менш ад *таго, што мы вымяраем*, чым ад таго, што мы будуем.

Коратка

Упэўненыя ілжывыя адказы моўных мадэляў маюць дзве крыніцы. Першая статыстычная: калі факт не мае заканамернасці, якую можна вывучыць, мадэль, навучаная імітаваць мову, часам будзе памыляцца, і гэтую ніжнюю мяжу можна ацаніць — напрыклад, па долі фактаў, якія сустракаюцца ў навучанні толькі адзін раз. Другая крыніца — стымулы: амаль кожны бенчмарк, паводле якога ранжыруюць мадэлі, ацэньвае «я не ведаю» гэтак жа, як няправільны адказ, таму ўгадваць заўсёды выгадней — аж да сітуацыі, калі мадэль, якая памыляецца ў трох чвэрцях выпадкаў, можа абагнаць больш сумленную мадэль, што ўстрымліваецца ад адказу. Прапанова аўтараў — не яшчэ адзін тэст на галюцынацыі, а **яўныя крытэрыі ацэнкі** (*open rubrics*): запісваць правіла ацэньвання проста ў пытанні. У дэманстрацыйным даследаванні на чатырох перадавых мадэлях гэта пераварочвае стымул, і метад памяншэння галюцынацый узнагароджваецца, а не караецца. Гэта тэарэтычная праца плюс агляд з невялікім, яўна некантраляваным эксперыментам, рэцэнзаваная ў *Nature*. Выпраўленне перспектыўнае, але яшчэ не паказана ў вялікім маштабе; самі галюцынацыі тут трактуюцца як не містычныя і не строга непазбежныя.

Праверка без прыкрас

Што паказвае артыкул: Матэматычную ніжнюю мяжу, пры якой некаторыя галюцынацыі ўзнікаюць ужо падчас папярэдняга навучання — для фактаў без патэрнаў прынамсі на ўзроўні singleton rate; агляд, паводле якога большасць вядучых бенчмаркаў не дае ніякага бонусу за «я не ведаю»; і дэманстрацыйнае даследаванне на чатырох мадэлях, дзе яўнае фармуляванне правілаў ацэнкі ў запыце — робіць метад памяншэння

галюцынацый выгадным там, дзе простая метрыка дакладнасці яго карала.

Што праўдападобна, але не даказана: Што ўкараненне яўных крытэрыяў ацэнкі ў асноўныя бенчмаркі істотна паменшыць галюцынацыі ў разгорнутых мадэлях. Падтрымліваючы эксперымент невялікі і наўпрост названы некантраляваным.

Чаго яна не паказвае: Што галюцынацыі непазбежныя — праца сцвярджае супрацьлеглае; што ацэньванне з'яўляецца адзінай прычынай; што галюцынацыі можна цалкам ліквідаваць; што дэманстрацыйнае даследаванне ранжыруе чатыры мадэлі паміж сабой.

Галоўныя абмежаванні: Наўмысна спрошчаная мадэль «правільна/няправільна/я не ведаю», якую самі аўтары называюць «ілжывай трыхатаміяй»; невялікі ўручную адабраны агляд бенчмаркаў — дзесяць ацэнак; некантраляванае дэманстрацыйнае даследаванне — чатыры мадэлі, адзін метад змяншэння галюцынацый, адзін тэст, стандартныя налады; і аргумент, у значнай ступені падрыхтаваны даследчыкамі OpenAI, пра тое, як галіне варта ацэньваць мадэлі.

Якой упэўненасці заслугоўвае выснова для звычайнага чытача? Высокай — што галюцынацыі не з'яўляюцца ні містычнымі, ні строга непазбежнымі і што галоўныя бенчмаркі цяпер узнагароджваюць угадванне. Умеранай — што прапанаванае выпраўленне дапаможа: цяпер ёсць сапраўдны proof of concept, але яшчэ няма дэманстрацыі шырокай эфектыўнасці ў вялікім маштабе.

Крыніцы

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