The Clean Paper

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

Ինչու են լեզվային մոդելները hallucinate անում — և ինչու մեր scoring-ը օգնում է, որ դա շարունակվի

Լեզվային մոդելների confident falsehood-ները mysterious glitch չեն. դրանց մի մասը training-ի statistical հետևանք է, իսկ persistence-ը մասամբ գալիս է այն բանից, որ mainstream benchmarks-ը confident guess-ը ավելի լավ են գնահատում, քան ազնիվ «I don't know»-ը: Four frontier models-ի case study-ն ցույց է տալիս, որ scoring rules-ը prompt-ում բացահայտելը՝ «open rubrics», կարող է շրջել այդ incentive-ը:

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

Ինչու են մոդելները գուշակում, և ինչու մենք նրանց հենց դա ենք սովորեցրել

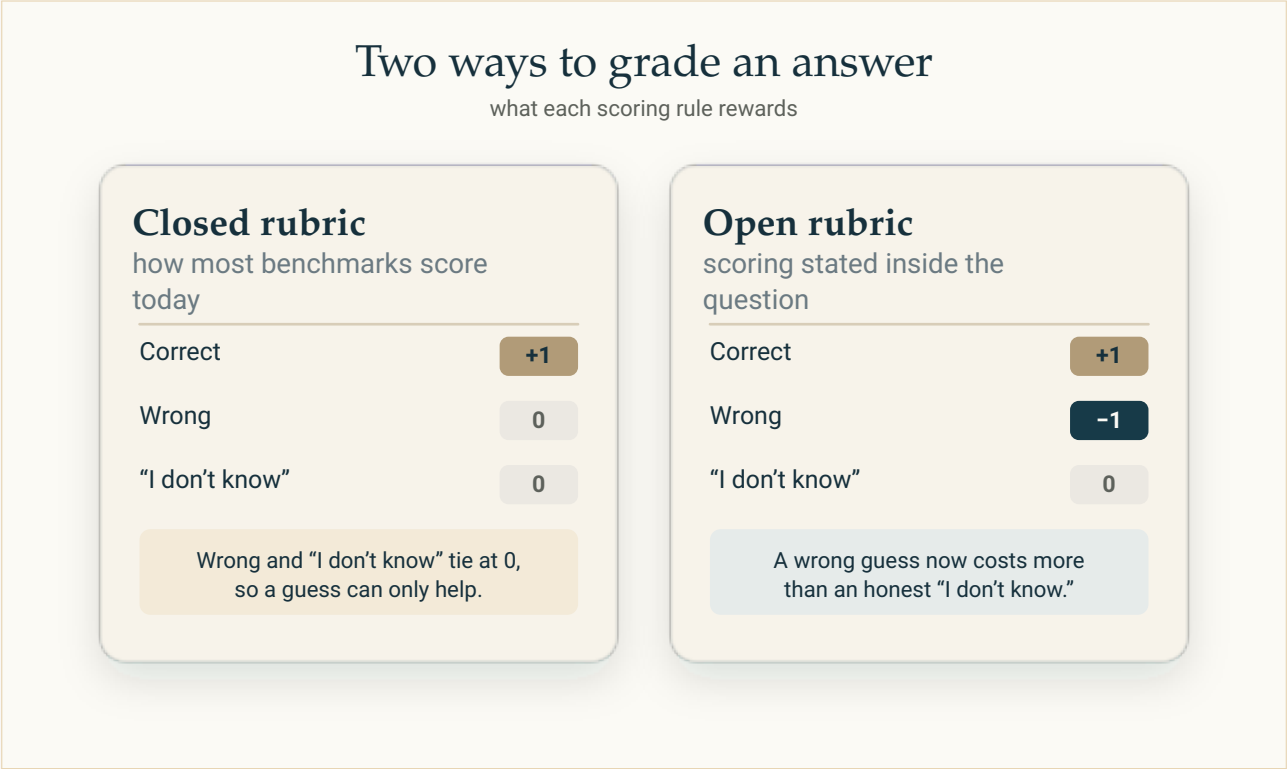
Մեծ լեզվային մոդելին հարցրեք անծանոթ մարդու ծննդյան օրը, և այն կարող է «մարտի 7» պատասխանել այնպիսի վստահությամբ, կարծես պատասխանը կարդում է քարտից: Հետո նույն հարցին երեք անգամ տալ երեք տարբեր ամսաթիվ՝ բոլորն էլ սխալ: Հեղինակներն հենց այսպիսի օրինակներ են բերում. առաջատար մոդելներին տալիս են սովորական փաստական հարց՝ որևէ մեկի ծննդյան օրը կամ հազվադեպ հապավման բացատրությունը, և յուրաքանչյուր մոդել վստահորեն հորինում է իր տարբերակը: Ոլորտում սա անվանում են *hallucination*՝ «հալյուցինացիա», բառ, որը խնդիրը ներկայացնում է որպես ընկալման խափանում: Հոդվածի առաջին կարևոր քայլն այն է, որ այդ առեղծվածային երանգը հանում է:

Սկսենք նրանից, թե ինչպես է կառուցվում մոդելը: Ամենամեծ՝ նախնական ուսուցման փուլում այն հսկայական քանակությամբ տեքստից սովորում է, թե ինչ տեսք ունի բնական և սահուն լեզուն: Հիմա վերցնենք մի փաստ, որի հետևում որևէ օրինաչափություն չկա՝ կոնկրետ մարդու ծննդյան օրը: Եթե այդ ամսաթիվը ուսուցման նյութում հանդիպել է ընդամենը մեկ անգամ կամ ընդհանրապես չի հանդիպել, օրինաչափություններ սովորող համակարգը բռնելու բան չունի. մոդելի տեսանկյունից պատասխանը կամայական է: Հեղինակները սա ձևակերպում են հին գաղափարի օգնությամբ, որը Ալան Թյուրինգը կիրառել էր բոլորովին այլ խնդրում: Եթե ուսուցման տվյալներում ծննդյան օրերի հինգերորդ մասը հանդիպում է միայն մեկ անգամ, ապա պետք է ակնկալել, որ մոդելը դրանցից առնվազն մոտ հինգերորդ մասում սխալվելու է: Ոչ թե որովհետև մոդելը «կոտրված» է, այլ որովհետև սովորելու համար բավարար տեղեկություն երբեք չի եղել: Նույն տրամաբանությամբ երկրների մայրաքաղաքների դեպքում մոդելներն հազվադեպ են սխալվում. այդ փաստերը տեքստերում անընդհատ կրկնվում են: Հեղինակներն ավելի ընդհանուր ձևով ցույց են տալիս, որ ճշմարիտ պնդումը հավանական, բայց կեղծ պնդումից տարբերակելն ինքնին դժվար խնդիր է, իսկ *միայն* ճշմարիտ պնդումներ արտադրելը առնվազն նույնքան դժվար է: Այսինքն՝ սխալի որոշակի ստորին սահման անխուսափելի է:

> **Հիմքում ընկած գաղափարը. ինչպես հաշվել այն, ինչ դեռ չեք տեսել**
>
> Սա հիմնվում է շատ գեղեցիկ գաղափարի վրա, որը լեզվային մոդելներից շատ ավելի հին է:
> **Պատկերացրեք գունավոր գնդիկներով պարկ**: Դուք չգիտեք, թե քանի գույն կա ներսում: Հանում եք 100 գնդիկ և հաշվում. կարմիր՝ 40, կապույտ՝ 25, կանաչ՝ 15, դեղին՝ 5, մանուշակագույն՝ 3, նարնջագույն՝ 2, իսկ հետո՝ **տասը տարբեր գույն, որոնցից յուրաքանչյուրը հանդիպում է ճիշտ մեկ անգամ**:
> Թյուրինգի հարցն այլ խնդրում այսպիսի բան էր. *որքա՞ն է հավանականությունը, որ հաջորդ գնդիկը կլինի մի գույն, որը մինչ այդ ընդհանրապես չեք տեսել*: Չտեսած գույները ուղղակի հաշվել չեք կարող: Բայց

> կարող եք հաշվել մեկ անգամ հանդիպած գույները՝ այսպես կոչված «միակ հանդիպումները»: **Գուդ-Թյուրինգի գնահատման** գաղափարն այն է, որ մեկ անգամ հանդիպած դիտարկումների բաժինը մոտավոր ցույց է տալիս, թե որքան հավանականություն է թաքնված դեռ չտեսած տարբերակների մեջ: Եթե 100 հանումներից 10-ը մեկանգամյա գույներ էին, ապա հաջորդ գնդիկի բոլորովին նոր գույն լինելու հավանականությունը մոտ **10 / 100 = 10%** է:
> Այդ մեկանգամյա գույները սխալներ չեն: Դրանք ձեր անտեղյակության չափումն են. եթե շատ բան հանդիպում է միայն մեկ անգամ, նմուշը հուշում է, որ աշխարհում դեռ շատ բան կա, որը պարզապես չեք տեսել:
> **Հիմա գույների փոխարեն պատկերացրեք ծննդյան օրեր**, իսկ պարկի փոխարեն՝ մոդելի ուսուցման տեքստերը: Ենթադրենք մոդելի տեսած ծննդյան օրերից **հինգից մեկը հանդիպել է միայն մեկ անգամ**: Նույն տրամաբանությամբ հավանականության մոտ մեկ հինգերորդը վերաբերում է փաստերի, որոնք մոդելը գործնականում չի սովորել: Իսկ ծննդյան օրը հնարավոր չէ «հաշվել» օրինաչափությունից: Հետևաբար մեկ անգամ կամ երբեք չտեսած ամսաթիվը մոդելի համար վատ տեղեկացված գուշակություն է, և մոտավորապես **հինգից մեկում** այն սխալվելու է: Այստեղ ավելի «խելացի» լինելը չի լուծում խնդիրը. տվյալներում սովորելու նյութ չի եղել:
> Փոքրացված ձևով սա ամբողջ փաստարկն է. մեկանգամյա դիտարկումների հաճախականությունը չափում է, թե տվյալ հավաքածուից աշխարհի որ մասն է դժվար կամ անհնար սովորել, և դա սխալների համար **ստորին սահման** է տալիս: Նույն պատճառով մայրաքաղաքների մասին հարցերում սխալը շատ հազվադեպ է. *Փարիզը* հանդիպում է անընդհատ, ուստի սովորելու նյութը շատ է:

Սա բացատրում է, թե որտեղից են գալիս հալյուցինացիաները: Բայց դեռ չի բացատրում, թե ինչու են դրանք *պահպանվում* նույնիսկ այն հետագա ուսուցումից հետո, որի նպատակը մոդելներին ավելի օգտակար ու ազնիվ դարձնելն է: Հոդվածի համեմատությունը այստեղ բավական դիպուկ է: Պատկերացրեք քննության ժամանակ մի ուսանողի, որը պատասխանը չգիտի: Եթե դատարկ պատասխանը զրո միավոր է ստանում, իսկ գուշակությունը երբեմն կարող է մեկ միավոր բերել, գնահատականը բարձրացնելու լավագույն ռազմավարությունը գուշակելն է՝ վստահ ու կոնկրետ, ոչ թե ասել «չգիտեմ»: Ուսանողները դա սովորում են: Մոդելներն էլ են սովորում, որովհետև մենք նրանց նույն կերպ ենք գնահատում: Հեղինակները դիտարկել են այն չափորոշիչները և մրցատախտակները, որոնց վրա ոլորտը իրականում մրցում է, և գտել, որ դրանց գրեթե բոլորը «չգիտեմ» պատասխանը գնահատում են ճիշտ նույն կերպ, ինչպես սխալ պատասխանը՝ զրո: Այդ կանոններով միշտ գուշակող մոդելը առավելություն ունի նույնքան կարող, բայց իր անորոշությունը ազնվորեն նշող մոդելի նկատմամբ: Այսինքն՝ բավական բառացի իմաստով մենք ինքներս ենք գնահատման համակարգով խրախուսում գուշակելը:



Չափորոշիչները կարող են գուշակելը դարձնել ռացիոնալ ռազմավարություն: Սովորական գնահատման դեպքում (ձախ) սխալ պատասխանը և ազնիվ «չգիտեմ»-ը երկուսն էլ զրո միավոր են ստանում, հետևաբար գուշակելը միայն օգուտ տալ կարող է: Եթե հարցի մեջ բացահայտ նշվում է գնահատման կանոնը և սխալ պատասխանի համար տուգանք կա (աջ), անորոշության դեպքում պատասխանելուց հրաժարվելը կարող է ավելի շահավետ լինել: Սա փոխում է այն, ինչ խրախուսում է թեստը, բայց ինքնին հայտուցինացիան չի վերացնում:

Հենց սա է հոդվածի ամենաարժեքավոր մասը, որովհետև այն հակադրվում է սովորական վերնագրերին: Հալյուցինացիան հաճախ ներկայացվում է որպես տեխնոլոգիայի գրեթե խորհրդավոր և անխուսափելի սահման: Հեղինակները վիճարկում են երկու գաղափարն էլ: Նախնական ուսուցման սխալի ստորին սահմանը առեղծված չէ. դա սովորական վիճակագրական սխալ է, որի հետ մեքենայական ուսուցումը տասնամյակներ շարունակ գործ ունի: Իսկ հետագա հալյուցինացիաների շարունակությունը ամբողջովին անխուսափելի չէ. դրա մի մասը մեր կառուցած խրախուսման համակարգի արդյունքն է, որը կարելի է փոխել: Համակարգը, որը անորոշության դեպքում պարզապես հրաժարվում է պատասխանելուց, կարող է խուսափել վստահ կեղծ պատասխաններից: Իրական մոդելները հաճախ այդպես չեն վարվում, որովհետև մեր մրցատախտակները հրաժարումը պատժում են:

Ի՞նչ արեցին հեղինակները

Հոդվածը երեք մաս ունի: Առաջինը մաթեմատիկական փաստարկ է, ըստ որի նախնական ուսուցման ընթացքում որոշակի քանակի հալյուցինացիաներ վիճակագրորեն

անխուսափելի են. հեղինակները ցույց են տալիս, որ «արտադրել միայն վավեր տեքստ» խնդիրը առնվազն նույնքան դժվար է, որքան երկընտրանքային «այս պնդումը վավե՞ր է» դասակարգման խնդիրը: Երկրորդ մասը, տասը ազդեցիկ չափորոշիչների ուսումնասիրության հիման վրա, պնդում է, որ սովորական ճշգրտության չափումները գուշակելը խրախուսում են, իսկ պատասխանելուց հրաժարվելը՝ ոչ: Երրորդում առաջարկվում և փորձարկվում է լուծման գաղափար՝ **բաց գնահատման կանոններ** (*open rubrics*): Գնահատման կանոնը գրվում է հենց հարցի մեջ, օրինակ՝ «ճիշտ պատասխանը +1 է, սխալը՝ −1, ուստի եթե վստահությունը 50%-ից ցածր է՝ ավելի լավ է չպատասխանել»: Այսպես մոդելը նախապես գիտի, թե երբ է ազնվությունը գնահատվելու: Հեղինակները դա փորձարկում են չորս առաջատար մոդելի՝ Google Gemini 3 Pro, OpenAI GPT-5, xAI Grok 4 և Anthropic Claude Opus 4.5-ի վրա՝ օգտագործելով SimpleQA-ի 4 326 փաստական հարցերը: Նրանք բացահայտ նշում են, որ սա ցուցադրական օրինակ է, ոչ մոդելների վերահսկվող համեմատություն. օգտագործվել են լռելյայն կարգավորումներ, առանց հատուկ հարմարեցման և ծախսերի հավասարեցման:

Ի՞նչ գտան

- **Նախնական ուսուցումն անխուսափելիորեն որոշ սխալ է թողնում**: Վստահ կեղծ պնդումներ արտադրելու հաճախականությունն ունի ստորին սահման, որը կապված է լավագույն «այս պնդումը վավե՞ր է» դասակարգչի սխալի հետ: Օրինաչափություն չունեցող փաստերի դեպքում այդ սահմանը առնվազն մեկանգամյա հանդիպումների հաճախականությունն է՝ այն փաստերի բաժինը, որոնք ուսուցման նյութում հանդիպում են միայն մեկ անգամ: Այսինքն՝ նույնիսկ կատարյալ մաքուր տվյալների դեպքում որոշ հայցուցինացիաներ անխուսափելի են:
- **Գնահատումը կոնկրետ կերպով խրախուսում է գուշակելը**: Սովորական ճիշտ/սխալ գնահատման դեպքում երբեք չհրաժարվելը օպտիմալ ռազմավարություն է, և հեղինակների ուսումնասիրած հայտնի չափորոշիչների մեծ մասը «չգիտեմ»-ը պարզապես սխալ է համարում: Վառ օրինակ՝ SimpleQA-ում մաքուր ճշգրտությունը փոքր առավելություն է տալիս OpenAI o4-mini-ին, որը պատասխանում է գրեթե ամեն ինչի և սխալվում է ավելի քան երեք քառորդ դեպքերում, GPT-5-mini-ի նկատմամբ, որը շատ ավելի քիչ է սխալվում, որովհետև անորոշության դեպքում հաճախ հրաժարվում է պատասխանել: Այսինքն՝ ավելի անկանխատեսելի մոդելը մրցատախտակում կարող է ավելի լավ երևալ:
- **Բաց գնահատման կանոնները փոխում են խրախուսումը**: Հեղինակները փորձարկում են պարզ հակահայցուցինացիոն մեթոդ՝ մոդելին տալ նույն հարցը երկու անգամ և, եթե պատասխանները չեն համընկնում, հրաժարվել վերջնական պատասխանից: Սովորական ճշգրտության դեպքում այս մեթոդը նվազեցնում է սխալները, բայց նաև նվազեցնում է «accuracy»-ն, ուստի չափանիշը կարծես պատժում է այն: Բաց գնահատման կանոններով նույն մեթոդը չորս մոդելների համար էլ ավելի լավ արդյունք է տալիս տարբեր տուգանքների պայմաններում: GPT-5-mini-ն էլ, որը սովորական ճշգրտությամբ տուժում էր հրաժարվելու համար,

բաց կանոններով անցնում է o4-mini-ից առաջ (յուրաքանչյուր մոդելի համար n = 4 326 հարց)։

Ի՞նչ է սա հավանաբար նշանակում

Հայցուցինացիան նվազեցնելը հիմնականում նոր՝ հատուկ հայցուցինացիաներ որսացող թեստեր հորինելու խնդիր չէ։ Ավելի կարևոր է փոխել այն, թե հիմնական չափորոշիչները ինչպես են գնահատում անորոշությունը, որպեսզի «չգիտեմ» ասելը այլևս չպատժվի։ Քանի դեռ մրցատախտակը չի փոխվում, հայցուցինացիան նվազեցնելը հաճախ մոդելին «ճշգրտության միավորներ» է արժենալու և հետևաբար քիչ խրախուսված կլինի։ Այդ պատճառով հեղինակները խնդիրը կոչում են «սոցիոտեխնիկական մի մասը ավելի լավ չափման համակարգ է, մյուս մասը՝ ազդեցիկ մրցատախտակներին այն ընդունել տալը։

Ի՞նչ սա *չի* ապացուցում

- Սա **չի ցույց տալիս, որ բաց գնահատման կանոնները իրական համակարգերում վերացնում են հայցուցինացիան**։ Փորձը փոքր և դիտավորյալ **չվերահսկվող** օրինակ է՝ չորս մոդել լռելյայն կարգավորումներով, մեկ հակահայցուցինացիոն մեթոդ և մեկ փաստական հարցերի թեստ։ Նպատակը խրախուսման շրջադարձը ցույց տալն է, ոչ մոդելներին դասակարգելը կամ ընդհանուր արդյունավետություն ապացուցելը։
- Սա **չի պնդում, որ գնահատումը միակ պատճառն է**։ Ուսուցման տվյալների սխալները, իսկապես դժվար խնդիրները և անծանոթ հարցումները մնում են առանձին աղբյուրներ։
- Սա **չի աջակցում այն տարածված պնդմանը, թե հայցուցինացիան անխուսափելի է**։ Հեղինակներն, ընդհակառակը, պնդում են, որ համակարգը, որը պատասխանում է միայն ստուգելի հարցերին և մնացած դեպքերում ասում «չգիտեմ», վստահ կեղծ պատասխաններ չի տա։
- Սա **չի վերացնում նախնական ուսուցման սխալի ստորին սահմանը**։ Այն բացատրում և սահմանափակում է այդ սխալը, ընդ որում խոսքը վստահ փաստական սխալների, ոչ մոդելի ամբողջ վարքի մասին է։
- Սա **չի ցույց տալիս, որ բաց գնահատման կանոններն ինքնուրույն բավարար են**։ Դրանք փոխում են գնահատման խրախուսումները, բայց չեն փոխարինում որոնմանը, գործիքների օգտագործմանը կամ ավելի լավ կարգաբերված մոդելներին։

Որքա՞ն ուժեղ է ապացույցը

- Հոդվածի հիմքը **մաթեմատիկան** է՝ ֆորմալ ստորին սահմաններ, ոչ փորձարարական չափումներ: Տեսական փաստարկն իր ենթադրությունների ներքո ինքնուրույն է:
- Այն հիմնվում է խնդրի **դիտավորյալ պարզեցված մոդելների** վրա: Հեղինակներն իրենք նշում են «կեղծ եռաբաժանումը», երբ ամեն պատասխան դասվում է ճիշտ, սխալ կամ «չգիտեմ» խմբերից մեկում, և ամենամաքուր սահմանների համար օգտագործվող «կամայական փաստերի» իդեալականացված միջավայրը:
- Չափորոշիչների ուսումնասիրությունը **փոքր, ընտրված նմուշ** է՝ տասը ազդեցիկ գնահատում, ոչ ամբողջ ոլորտի սպառիչ աուդիտ:
- Փորձնական օրինակը **իրական է, բայց սահմանափակ**՝ չորս առաջատար մոդել, մեկ մեթոդ, միայն SimpleQA, լռելյայն կարգավորումներ և բացահայտ նշված «ոչ վերահսկվող գնահատում»: Սա խրախուսման գաղափարի ապացույց է, ոչ նոր մրցատախտակ:
- Կարևոր է նաև նշել տեսանկյունը. չորս հեղինակներից երեքը աշխատում են կամ աշխատել են **OpenAI-ում**, իսկ հոդվածը առաջարկում է փոխել ամբողջ ոլորտի գնահատման մեթոդները: Սա շահագրգիռ կողմի լավ փաստարկված դիրքորոշում է, ոչ չեզոք արտաքին գրախոսություն. արժե հաշվի առնել, բայց ոչ դրա համար մերժել: Հոդվածը, ի դեպ, նույն քննադատությունը կիրառում է նաև սեփական o4-mini և GPT-5-mini մոդելների նկատմամբ:

Ինչո՞ւ է սա կարևոր

Աշխատանքը վերաձևակերպում է չափազանցված թեմա: «Հալյուցինացիան» հաճախ ներկայացվում է կամ որպես խորհրդավոր խափանում, կամ որպես անշարժ տեխնիկական պատ: Այս հոդվածը այն դարձնում է ավելի սովորական և մասամբ մեր իսկ ստեղծած խնդիր՝ վիճակագրական ստորին սահման, որը կարող ենք հասկանալ, գումարած խրախուսման համակարգ, որը մենք ենք ընտրել: Ավելի լայն դասը պարզ է. հուսալիության հետագա առաջընթացը կարող է կախված լինել ոչ միայն նրանից, թե ինչ ենք կառուցում, այլ նաև նրանից, թե **ինչ ենք չափում**:

Կարճ ամփոփում

Լեզվային մոդելների վստահ կեղծ պատասխանները երկու հիմնական աղբյուր ունեն: Առաջինը վիճակագրական է. եթե որևէ փաստի մեջ սովորելի օրինաչափություն չկա, լեզուն ընդօրինակելու համար ուսուցված մոդելը երբեմն այն սխալ է գուշակելու, և այդ սխալի ստորին սահմանը կարելի է գնահատել, օրինակ, ըստ այն բանի, թե քանի փաստ է ուսուցման տվյալներում հանդիպում միայն մեկ անգամ: Երկրորդ աղբյուրը խրախուսումն է. մոդելների գրեթե բոլոր հայտնի չափորոշիչներում «չգիտեմ»-ը նույնքան վատ է գնահատվում, որքան սխալ պատասխանը, հետևաբար

գուշակելը միշտ ավելի շահավետ է: Հեղինակների առաջարկը նոր հայցուցինացիոն թեստ չէ, այլ **բաց գնահատման կանոններ**՝ գնահատման ձևը հենց հարցի մեջ գրել: Չորս առաջատար մոդելների փոքր փորձնական օրինակում դա փոխում է խրախուսումը, և հայցուցինացիան նվազեցնող մեթոդը սկսում է պարգևատրվել, ոչ պատժվել: Սա տեսություն, չափորոշիչների ուսումնասիրություն և փոքր, բացահայտ չվերահսկվող փորձ միավորող, *Nature*-ում գրախոսված աշխատանք է: Առաջարկը խոստումնալից է, բայց դեռ լայն մասշտաբով ապացուցված չէ: Հայցուցինացիաները ներկայացվում են ոչ որպես խորհրդավոր, ոչ էլ որպես խիստ անխուսափելի երևույթ:

Առանց չափազանցության

Ի՞նչ է ցույց տալիս աշխատանքը: Մաթեմատիկական ստորին սահման, որի պատճառով նախնական ուսուցման ժամանակ որոշ վստահ փաստական սխալներ անխուսափելի են, «մեկանգամյա հանդիպումների» հաճախականության գաղափարը օրինաչափություն չունեցող փաստերի համար, ուսումնասիրություն, ըստ որի առաջատար չափորոշիչների մեծ մասը «չգիտեմ»-ին միավոր չի տալիս, և չորս մոդելի փորձնական օրինակ, որտեղ գնահատման կանոնը հարցի մեջ բացահայտ գրելը հայցուցինացիա նվազեցնող մեթոդը դարձնում է շահավետ այն դեպքում, երբ սովորական ճշգրտությունը այն պատժում էր:

Ի՞նչն է հավանական, բայց ապացուցված չէ: Որ բաց գնահատման կանոնները, եթե դառնան հիմնական չափորոշիչների մաս, զգալիորեն կնվազեցնեն իրական կիրառվող մոդելների հայցուցինացիաները: Աջակցող փորձը փոքր է և բացահայտ չվերահսկվող:

Ի՞նչ չի ցույց տալիս: Որ հայցուցինացիաներն անխուսափելի են՝ հոդվածը հակառակն է պնդում, որ գնահատումը միակ պատճառն է, որ հայցուցինացիան կարելի է ամբողջությամբ վերացնել, կամ որ փորձնական օրինակը չորս մոդելների համեմատական վարկանիշ է տալիս:

Հիմնական սահմանափակումները: Դիտավորյալ պարզեցված ճիշտ/սխալ/«չգիտեմ» մոդել, տասը չափորոշիչից կազմված փոքր ընտրված նմուշ, չորս մոդելով, մեկ մեթոդով և մեկ թեստով չվերահսկվող փորձ, և OpenAI-ի հետ կապ ունեցող հեղինակների փաստարկ այն մասին, թե ինչպես պետք է ոլորտը գնահատի մոդելները:

Որքա՞ն վստահություն պետք է ունենա ընդհանուր ընթերցողը: Բարձր՝ որ հայցուցինացիան ոչ խորհրդավոր է, ոչ էլ խիստ անխուսափելի, և որ հիմնական չափորոշիչներն այսօր իսկապես խրախուսում են գուշակելը: Միջին՝ առաջարկվող լուծման լայն արդյունավետության նկատմամբ. կա իրական գաղափարի ապացույց, բայց դեռ չկա լայն մասշտաբով ցուցադրում:

Աղբյուրներ

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