The Clean Paper

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

Miért hallucinálnak a nyelvi modellek — és miért tartja fenn ezt az, ahogyan pontozzuk őket

A magabiztos hamis állítások, amelyeket „hallucinációnak” nevezünk, nem rejtélyes hibák: egy részük a tanítás statisztikai melléktermése, és azért maradnak fenn, mert a fősodorbeli benchmarkok jobban jutalmazzák a magabiztos tippet, mint az őszinte „nem tudom” választ. Négy élvonalbeli modellen végzett esettanulmány szerint a pontozási szabályok promptban való közlése („nyílt pontozási szabályok”) megfordítja ezt az ösztönzőt.

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

Miért tippelnek a modellek — és miért mi tanítottuk meg őket erre

Kérdezzünk meg egy nagy nyelvi modellt egy idegen ember születésnapjáról, és könnyen azt válaszolhatja, hogy „március 7.”, olyan nyugodt magabiztossággal, mintha egy kártyáról olvasná — miközben téved, akár három egymást követő alkalommal három különböző dátummal. A szerzők pontosan ilyen példákat hoznak: vezető modelleknek egyszerű ténykérdést tesznek fel — valakinek a születésnapját vagy egy kevéssé ismert betűszó feloldását —, és mindegyik modell magabiztosan kitalál egy másik választ, amelyek közül egyik sem helyes. Az iparág ezt *hallucinációnak* nevezi, amitől úgy hangzik, mintha az észlelés hibájáról lenne szó. A tanulmány első lépése éppen az, hogy eltünteti belőle a misztikumot.

Induljunk onnan, hogyan készül egy modell. A tanítás első és legnagyobb szakaszában lényegében azt tanulja meg hatalmas mennyiségű szöveg olvasásával, hogyan néz ki a folyékony nyelv. Most vegyünk egy olyan tényt, amely mögött nincs felismerhető minta — például egy konkrét ember születésnapját. Ha ez a dátum egyszer vagy egyáltalán nem szerepelt a tanító szövegben, egy mintázatokat tanuló rendszernek nincs mibe kapaszkodnia: a modell szempontjából a válasz önkényes. A szerzők ezt egy régi ötlet kölcsönzésével teszik precízzé — Alan Turing egy másik problémára adott gondolatával. Ha például minden ötödik születésnap csak egyszer jelenik meg az adatokban, akkor várható, hogy a modell legalább minden ötödiket elrontja — nem azért, mert hibás, hanem mert soha nem volt ott elegendő információ, amelyből megtanulhatta volna. (Ugyanezen logika alapján az országok fővárosait a modellek szinte soha nem rontják el: ezek folyamatosan előfordulnak.) A szerzők gondosan amellett érvelnek, hogy egy igaz állítást megkülönböztetni egy hihető hamistól önmagában nehéz probléma, és *csak* igaz állításokat generálni legalább ilyen nehéz. Van tehát egy beépített hibaküszöb.

Az alapötlet: hogyan számoljuk meg azt, amit még nem láttunk?

Ez egy valóban ötletes gondolatra épül — régebbire, mint a nyelvi modellek, és érdemes rendesen megismerni.

Kezdjünk egy zsák színes golyóval. Nem tudjuk, hányféle szín van benne. Egymás után kihúzunk 100-at és megszámoljuk őket: piros 40, kék 25, zöld 15, sárga 5, lila 3, narancssárga 2 — majd **tíz különböző szín, amelyek mind pontosan egyszer fordulnak elő.**

A kérdés, amellyel Turing valójában szembesült egy egészen más problémában, így szólt: *mekkora az esélye annak, hogy a következő golyó olyan színű lesz, amelyet eddig egyáltalán nem láttunk?* Azt, amit soha nem húztunk ki, nem tudjuk közvetlenül megszámolni — de meg tudjuk számolni azokat a színeket, amelyek pontosan egyszer jelentek meg, az úgynevezett **singletonokat**. A **Good–Turing-becslés** nevű trükk szerint az egyszer előforduló elemek aránya megbecsüli azt a valószínűségi tömeget, amely még a soha nem látott színekben rejtőzik. A száz húzásból tíz olyan szín volt, amely csak egyszer fordult elő, ezért annak esélye, hogy a következő golyó teljesen új színű lesz, körülbelül **10 / 100 = 10%.**

Az egyszer látott színek nem *hibák*. A saját tudatlanságunk mérőeszközei: ha sok szín csak egyszer jelenik meg a mintában, a minta ezzel azt jelzi, hogy a világban valószínűleg még több olyan szín van, amelyet egyszerűen nem húztunk ki.

Most cseréljük le a színeket születésnapokra, a zsákot pedig a modell tanító szövegére.

> Tegyük fel, hogy a látott születésnapok között **minden ötödik pontosan egyszer szerepel.** Ugyanaz a trükk: a valószínűség körülbelül egyötöde olyan születésnapokban van, amelyeket a modell gyakorlatilag soha nem látott — és egy születésnap mögött nincs minta, amelyre támaszkodhatna (nem lehet *kikövetkeztetni* valakinek a születésnapját). Egy egyszer vagy soha nem látott dátum tehát olyan érme, amelynek súlyozását a modell nem ismeri, és nagyjából **minden ötödik** esetben tévedni fog. Ezt semmilyen ügyesség nem javítja ki: nem volt ott megtanulható információ.
> Ez az egész érv kicsiben: a singletonok aránya azt méri, hogy *ebből az adathalmazból* a világ mekkora része tanulhatatlan, és ez **alsó korlátot** ad a hibákra. Azt is megmagyarázza, miért téveszt el egy modell alig valaha fővárost — *Párizs* állandóan előfordul, a singletonaránya közel nulla, ezért rengeteg adatból tanulható meg.

Ez megmagyarázza, honnan jönnek a hallucinációk. Azt azonban még nem, miért *maradnak meg* — miért blöffölnek a modellek a későbbi, segítőkészségre és őszinteségre irányuló tanítás után is ahelyett, hogy beismernék a bizonytalanságot. Itt a tanulmány hasonlata szinte kellemetlenül találó. Képzeljünk el egy vizsgázó diákot, aki nem tudja a választ. Ha az üresen hagyott kérdésért nulla pont jár, a tippért viszont akár egy is, akkor a jegyet maximalizáló stratégia a tippelés — magabiztosan, konkrétan, soha nem azt mondva, hogy „nem vagyok biztos benne”. A diákok megtanulják ezt. Úgy tűnik, a modellek is — mert ugyanígy osztályozzuk őket. A szerzők végignézték azokat a benchmarkokat, amelyekben a terület ténylegesen versenyez, és amelyek ranglistáin a modelleket feljebb próbálják tolni. Azt találták, hogy szinte mindegyik pontosan ugyanannyi pontot ad a „nem tudom” válaszra, mint a hibásra: nullát. Ilyen szabály mellett egy mindig tippelő modell jobb pontszámot ér el, mint egy minden másban azonos modell, amely őszintén jelzi a bizonytalanságát. Meglehetősen szó szerinti értelemben a pontozással tanítjuk bele őket ebbe a viselkedésbe.

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.”

A benchmarkok racionálissá tehetik a tippelést. A legtöbb benchmark pontozásában (balra) a hibás válasz és az őszinte „nem tudom" egyaránt nulla pontot ér — ezért egy tipp csak javíthat az eredményen. Ha a szabályokat magában a kérdésben közöljük, és a hibás válaszért büntetés jár (jobbra), bizonytalanság esetén a válasz megtagadása válhat jobb lépéssé. Ez megváltoztatja, mit jutalmaz a teszt; önmagában nem oldja meg a hallucinációt.

Original diagram — The Clean Paper CC BY 4.0

Ezt a részt érdemes megtartani, mert szembemegy a szokásos címsorral. A hallucinációt gyakran a technológia elkerülhetetlen, szinte misztikus korlátjaként adják el. A tanulmány mindkét részt vitatja. Az előtanítási hibapadló nem rejtély — hétköznapi statisztikai hiba, amilyet a gépi tanulás évtizedek óta ismer. A fennmaradása pedig nem elkerülhetetlen — részben olyan ösztönző, amelyet mi építettünk be, és meg is változtathatnánk. Egy rendszer, amely egyszerűen nem válaszolna, amikor bizonytalan, egyáltalán nem hallucinálna; azért nem így viselkednek a bevetett modellek, mert a pontozási rendszereink büntetik a visszautasítást.

Mit csináltak a szerzők?

A tanulmány három részből áll. Először matematikai érvet ad arra, hogy az előtanítás során bizonyos mennyiségű hallucináció statisztikailag kikényszerített: megmutatja, hogy „csak érvényes szöveget generálni" legalább olyan nehéz, mint a bináris „érvényes-e ez az állítás?" osztályozási probléma. Másodszor — tíz nagy hatású benchmark áttekintésével alátámasztva — amellett érvel, hogy a fősodorbeli, pontosságközpontú metrikák a tippelést jutalmazzák a válasz megtagadásával szemben. Harmadszor egy javítást javasol, majd esettanulmányban teszteli: **nyílt pontozási szabályú** (*open-rubric*) értékeléseket, amelyeknél a pontozást magában a kérdésben közlik (például: „a helyes válasz 1 pont, a hibás −1, ezért ha 50%-nál kevésbé vagy biztos, inkább ne válaszolj"), így a modell tudhatja, mikor jutalmazzák az őszinteséget. Ezt négy élvonalbeli modellen — a Google Gemini 3 Pro, az OpenAI GPT-5, az xAI Grok 4 és az Anthropic Claude Opus 4.5 modelljén — próbálják ki a SimpleQA 4,326 ténykérdésével. Kifejezetten hangsúlyozzák, hogy az esettanulmány illusztratív, „nem kontrollált modellek közötti értékelés" (alapbeállítások, nincs finomhangolás, nincs költségnormalizálás).

Mit találtak?

- **Az előtanítás bizonyos hibamennyiséget kikényszerít.** Annak arányát, hogy egy modell magabiztos hamis állításokat bocsát ki, alulról korlátozza a belőle felépíthető legjobb „érvényes-e ez az állítás?" osztályozó hibaaránya — nagyjából annak kétszerese. Olyan tényeknél, amelyek mögött nincs tanulható minta, ez az alsó korlát legalább a *singletonarány*, vagyis a tanítás során pontosan egyszer megjelenő tények aránya. Bizonyos mennyiségű hallucináció még tökéletesen tiszta adatok mellett is elkerülhetetlen.
- **A pontozás konkrétan jutalmazza a tippelést.** Szokásos helyes/hibás pontozás mellett az optimális stratégia az, hogy a modell soha ne tartózkodjon a válaszadástól; a szerzők áttekintése szerint a népszerű benchmarkok túlnyomó többsége a „nem tudom" választ egyszerűen hibásként

pontozza. Látványos példa a saját modelljeik közül: a SimpleQA teszten a nyers pontosság kissé *előnyben részesíti* az OpenAI o4-mini modelljét — amely szinte mindenre válaszol, és az esetek több mint háromnegyedében téved — a GPT-5-minihez képest, amely sokkal kevesebb hibát követ el, mert bizonytalanság esetén inkább tartózkodik. A vakmerőbb modell jobban néz ki a ranglistán.
- **A nyílt pontozási szabályok megfordítják az ösztönzőt — ebben az esettanulmányban.** Egy egyszerű hallucinációcsökkentő eljárást tesztelnek: a modell válaszoljon kétszer, és ha a két válasz eltér, inkább tartózkodjon. Szokásos pontosság mellett ez csökkenti a hibák számát, *de a pontosságot is csökkenti* — így a metrika az alkalmazása ellen ösztönöz. Nyílt pontozás mellett ugyanaz az eljárás mind a négy modellnél jobb eredményt ad a büntetések többféle értéke mellett; a GPT-5-mini pedig — amelyet a nyers pontosság azért büntetett, mert bizonytalanság esetén tartózkodott — megelőzi az o4-minit, amikor a pontozási szabályt nyíltan közlik (modellként n = 4,326 kérdés).

Mit jelent ez valószínűleg?

A hallucináció csökkentése nagyrészt nem újabb, kifejezetten hallucinációra tervezett tesztek kitalálásáról szól. Inkább arról, hogy megváltoztassuk, hogyan pontozzák a fősodorbeli benchmarkok a bizonytalanságot, hogy a „nem tudom” beismerése ne legyen többé büntetve. Amíg a ranglista nem változik, a hallucináció csökkentése továbbra is pontossági pontokba kerül majd a modelleknek, ezért az értékelési rendszer továbbra is ellene ösztönöz — ezért nevezik a szerzők a problémát „szocio-technikai” jellegűnek: részben jobb metrikára van szükség, részben arra, hogy a nagy hatású ranglisták valóban át is vegyék.

Mit nem bizonyít mindez?

- **Nem** mutatja, hogy a nyílt pontozás a való világban megoldja a hallucinációt. Az alátámasztó kísérlet kicsi, szándékosan **nem kontrollált** esettanulmány — négy modell alapbeállításokkal, egy kiválasztott mérséklési módszer, egy ténykérdéses teszt —, amelynek célja az *ösztönző megfordulásának* bemutatása, nem a modellek rangsorolása vagy az általános hatékonyság bizonyítása.
- **Nem** állítja, hogy a pontozás az *egyetlen* ok. A tanító adatok hibái, a valóban nehéz problémák és az ismeretlen promptok külön hibaforrások maradnak.
- **Nem** támasztja alá azt a népszerű állítást, hogy a hallucinációk elkerülhetetlenek. A szerzők ennek éppen az ellenkezője mellett érvelnek: egy rendszer, amely csak ellenőrizhető kérdésekre válaszolna, máskor pedig azt mondaná, hogy „nem tudom”, soha nem hallucinálna.
- **Nem** tünteti el az előtanítási hibapadlót — megmagyarázza és korlátozza, a korlát pedig a magabiztos ténybeli hibákra vonatkozik, nem a modell teljes viselkedésére.
- **Nem** mutatja, hogy a nyílt pontozási szabályok önmagukban *elegendők*. Azt változtatják meg, mit jutalmaz az értékelés; nem helyettesítik a visszakeresést, az eszközhasználatot vagy a jobban kalibrált modelleket.

Mennyire erős a bizonyíték?

- A mag **matematika** — formális alsó korlátok, nem mérések. Elméleti érvként a saját feltételei mellett konzisztens.
- A probléma **szándékosan leegyszerűsített modelljeire** épül; maguk a szerzők is kiemelik annak „hamis trichotómiáját", hogy minden választ helyesnek, hibásnak vagy „nem tudom"-nak tekintünk, valamint a legtisztább alsó korláthoz használt idealizált „önkényes tények" környezetét.
- A benchmarkok áttekintése **kicsi, válogatott minta** — tíz nagy hatású értékelés, nem teljes audit.
- Az esettanulmány **valós, de korlátozott**: négy élvonalbeli modell, egyetlen mérséklési eljárás, csak SimpleQA, alapbeállítások, és kifejezetten „nem kontrollált értékelés". Az ösztönzőről szóló érv proof of conceptje, nem benchmarkeredmény.
- Érdemes megnevezni a nézőpontot: a négy szerzőből hárman az **OpenAI** alkalmazottai vagy korábbi alkalmazottai, a tanulmány pedig amellett érvel, hogy a területnek meg kellene változtatnia a modellek értékelését. Ez egy érdekelt fél jól felépített álláspontja, nem semleges külső felülvizsgálat — mérlegelni kell, nem elutasítani. (A tanulmány javára szól, hogy a kritikát ugyanúgy saját modelljeire, az o4-minire és a GPT-5-minire is alkalmazza, mint másokra.)

Miért számít?

Más keretbe helyez egy erősen túlhájpolt problémát. A „hallucinációt" hajlamosak vagy kísérteties hibaként, vagy mozdíthatatlan falként bemutatni; ez a tanulmány hétköznapibbá és részben saját magunk által előidézetté teszi — van egy statisztikai hibapadló, amelyet ténylegesen megérthetünk, és erre rakódik rá egy általunk választott ösztönző. A tágabb tanulság csendesebb és hasznosabb: a megbízhatóság további javulása legalább annyira függhet attól, *mit mérünk*, mint attól, mit építünk.

Röviden

A nyelvi modellek magabiztos hamis válaszai két helyről származnak. Az első statisztikai: ha egy tény mögött nincs tanulható minta, a nyelv utánzására tanított modell időnként tévedni fog, és ennek hibapadlója megbecsülhető — például abból, hány tény jelenik meg pontosan egyszer a tanítás során. A második az ösztönzőrendszer: szinte minden benchmark, amelyen a modelleket rangsorolják, ugyanúgy pontozza a „nem tudom" választ, mint a hibásat, ezért a tippelés mindig nyer — olyannyira, hogy egy az esetek háromnegyedében hibázó modell jobb eredményt érhet el egy őszintébb, bizonytalanság esetén tartózkodó modellnél. A szerzők javaslata nem egy újabb hallucinációteszt, hanem a „nyílt pontozási szabály": a kérdésben közöljük a pontozást. Négy élvonalbeli modellen végzett esettanulmányban ez megfordítja az ösztönzőt, így a hallucinációt csökkentő módszert jutalmazza ahelyett, hogy büntetné. Ez elméleti és benchmark-felmérési tanulmány egy kicsi, kifejezetten nem kontrollált kísérlettel, szakértői bírálaton átesve a *Nature*-ben; a javítás ígéretes, de még nem mutatták meg, hogy nagy léptékben is működik, a hallucinációkról pedig a szerzők azt állítják, hogy sem nem misztikusak, sem nem szigorúan elkerülhetetlenek.

Tárgyilagos mérleg

Mit mutat a tanulmány? Matematikai alsó korlátot, amely mellett az előtanítás során bizonyos hallucináció kikényszerített — tanulható minta nélküli tényeknél legalább a „singletonarány”; egy felmérést, amely szerint a legtöbb vezető benchmark nem ad pontot a „nem tudom” válaszra; valamint egy négymodelles esettanulmányt, amelyben a pontozás promptban történő közlése („nyílt pontozási szabály”) nyertessé tesz egy hallucinációcsökkentő módszert ott, ahol a sima pontosság büntette.

Mi hihető, de nem bizonyított? Hogy a fősodorbeli benchmarkokba beépített nyílt pontozási szabályok érdemben csökkentenék a bevetett modellek hallucinációját. Az ezt alátámasztó kísérlet kicsi és kifejezetten nem kontrollált.

Mit nem mutat? Hogy a hallucinációk elkerülhetetlenek — éppen az ellenkezőjét állítja; hogy a pontozás az egyetlen ok; hogy a hallucináció teljesen megszüntethető; vagy hogy az esettanulmány rangsorolja egymáshoz képest a négy modellt.

Fő korlátok: Szándékosan leegyszerűsített helyes/hibás/„nem tudom” modell — a szerzők ezt „hamis trichotómiának” nevezik; kis, válogatott benchmarkminta, tíz értékeléssel; nem kontrollált esettanulmány négy modellel, egy mérséklési módszerrel, egy teszttel és alapbeállításokkal; valamint egy OpenAI-vezette érv arról, hogyan kellene a területnek modelleket értékelnie.

Mekkora bizalom indokolt egy általános olvasó részéről? Magas abban, hogy a hallucináció sem nem misztikus, sem nem szigorúan elkerülhetetlen, és hogy a fősodorbeli benchmarkok jelenleg jutalmazzák a tippelést. Közepes abban, hogy a javasolt javítás segít — már van valódi proof of concept, de még nincs demonstráció arra, hogy széles körben és nagy léptékben is működik.

Források

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