Vox Machinae: A tömegek bölcsessége gépi sokaságokban

Bartha Zoltán, PhD
Miskolci Egyetem
bartha.zoltan@ekon.me

Absztrakt

Galton klasszikus megfigyelése szerint a laikus becslések mediánja a valós érték jó közelítését adja. Jelen tanulmány ezt a kísérletet ismétli meg egy olyan sokaságon, amelynek tagjai nagy nyelvi modellek. Négy modellt kérdeztem le nyolc szerepkörben és két mintavételi beállítás mellett, kérdésenként 192 előrejelzést gyűjtve 47 valós Metaculus-kérdésre. A lekérdezés minden kérdésnél a lezárás napján, a kimenetel ismerete előtt történt, ami kizárja az adatszennyeződést. Az aggregálási hatás kimutatható: a bináris kérdéseken a sokaság mediánjának Brier-pontszáma 0,178, szemben az egyéni előrejelzések 0,211-es átlagával ($p = 0,004$). A hatás mértéke azonban jóval elmarad az eredetitől: a gépi medián az egyéni becslések 63,7 százalékát múlta felül, míg Galton sokaságán ez az arány mintegy 90 százalék volt. Az elmaradás oka a sokaság szerkezete: a sokféleség 95,6 százaléka egyetlen forrásból, a modellválasztásból származik, a modellek hibái pedig 0,90-es átlagos korrelációt mutatnak. Ebből következően a gépi sokaságnál az összetétel fontosabb a méretnél: a két gyengébb modell elhagyása javítja a pontosságot, miközben a sokaságméret-görbe már 16 előrejelzésnél ellaposodik. Az emberi közösségi előrejelzéshez mért hátrány lényegében a csaknem bizonyos kimenetelű kérdésekre korlátozódik.

Kulcsszavak

tömegek bölcsessége, aggregálás, nagy nyelvi modellek, előrejelzés, Brier-pontszám

Javasolt hivatkozás

Bartha, Z. (2026): Vox Machinae: A tömegek bölcsessége gépi sokaságokban. In: Bartha, Z. (szerk.): Statisztika+ Galton konferencia előadásai, 2026. szeptember 11. GEMI Műhelytanulmányok 2026/1, ISBN 978-963-358-455-2, ISSN 2939-5038, pp. 20–37.

Bevezetés

Francis Galton 1907-ben közölt megfigyelése, mely szerint egy vásári súlybecslő verseny tippjeinek mediánja szinte pontosan eltalálta a valós értéket, azóta a tömegek bölcsessége néven ismert, és a mögötte álló mechanizmust is jól ismerjük: ha a becslők hibái eltérő irányúak, az aggregálás során részben kioltják egymást.

Az utóbbi néhány évben ez a klasszikus kérdés új formában vetődött fel. A nagy nyelvi modellek képesek valószínűségi előrejelzést adni, és több vizsgálat is arra jutott, hogy bár egyetlen modell önmagában nem versenyképes az emberi előrejelzőkkel, több modell együttese már megközelítheti egy emberi sokaság teljesítményét. Kézenfekvő tehát a kérdés: működik-e a tömegek bölcsessége akkor is, ha a sokaság tagjai nem emberek, hanem nyelvi modellek.

A kérdés azért nem magától értetődő, mert egy gépi sokaság a klasszikus feltételek egyikét szerkezetéből adódóan sérti: a nyelvi modellek hasonló architekturális elveken nyugszanak és

részben átfedő tanítóadatokon tanultak, ezért ítéleteik nem függetlenek egymástól. Nyitott kérdés, hogy marad-e közöttük elegendő hibasokféleség ahhoz, hogy az aggregálás érdemi javulást hozzon.

Ebben a tanulmányban Galton kísérletét ismétlem meg egy mesterségesen előállított gépi sokaságon. Négy különböző fejlesztőtől származó nyelvi modellt kérdeztem le nyolc szerepkörben és két mintavételi beállítás mellett, kérdésenként három ismétlésben, így kérdésenként 192 előrejelzést gyűjtve. A kérdések 47 valós, a Metaculus platformon közzétett előrejelzési kérdésből állnak, amelyek 2026 nyarán zárultak le. A lekérdezés minden esetben a kérdés lezárási napján történt, vagyis akkor, amikor a kimenetel még nem volt ismert.

Az eredmények szerint az aggregálási hatás kimutatható, de az eredetinél lényegesen gyengébb, és a gyengeség oka a sokaság szerkezetében keresendő. A vizsgálat emellett azt is megmutatja, hogy a gépi sokaságnál az összetétel fontosabb a méretnél, és hogy a gépi előrejelzés hátránya az emberi sokasághoz képest lényegében egyetlen kérdéstípus miatt áll elő.

Az irodalomáttekintés a tömegek bölcsességének irodalmát és a nyelvi modellek előrejelzési képességét vizsgáló újabb kutatásokat tekinti át, majd megfogalmazza a kutatási rést. Ezt a módszertan, majd az eredmények bemutatása követi; a konklúzió a további kutatási irányokat is kijelöli.

Irodalomáttekintés

Galton megfigyelése és a tömegek bölcsessége

A kollektív előrejelzés kutatásának kiindulópontja egy 1906-os megfigyelés. Francis Galton a nyugat-angliai Plymouth állatkiállításán egy súlybecslő versenyt figyelt meg, ahol 787 látogató tippelt arra, mennyi lesz egy ökör tisztított súlya. Galtont eredetileg az foglalkoztatta, mennyire téved nagyot a laikus tömeg, és a beérkezett szelvényeket ennek igazolásának szándékával dolgozta fel. Az eredmény azonban pont az ellenkezőjét bizonyította: a tippek mediánja egy százalékon belül megközelítette a valós értéket, és pontosabb volt, mint a résztvevők túlnyomó többségének (beleértve a szakértő henteseket és állattenyésztőket) becslése (Galton, 1907).

A megfigyelés jelentősége az aggregálás szerepére mutat rá: a sokaság együttes ítélete olyan információt hordozott, amellyel a tagjai sem rendelkeztek. Galton a mediánt választotta összegző mértéknek, és a szélsőséges tippekkel szembeni robusztussága miatt érvelt mellette. Máig vitatott, hogy mi a legjobb összegzési megoldás; Galton mintájának érdekességeire a numerikus becsléseket tárgyaló résznél térek vissza.

A jelenség, amelyet ma a tömegek bölcsessége néven ismerünk, Surowiecki (2004) összefoglalója nyomán vált széles körben ismertté. Surowiecki négy feltételt fogalmazott meg, amelyek teljesülése esetén a sokaság ítélete megbízhatóan felülmúlja tagjaiét: a vélemények sokfélesége, az ítéletek függetlensége, a tudás decentralizáltsága és a megfelelő aggregálási mechanizmus megléte. Ez a négyes máig használható elemzési keret, a jelen vizsgálat szempontjából azonban két eleme, a sokféleség és a függetlenség szorul pontosításra.

Miért működik, és mit jelent valójában a sokféleség

Az aggregálási nyereség szokásos magyarázata a hibakioltás: ha az egyéni becslések hibái véletlenszerűek és eltérő irányúak, akkor átlagoláskor részben kioltják egymást, miközben a közös jelzés megmarad (Larrick & Soll, 2006). A kérdés az, mitől függ, mennyi hiba oltódik ki.

Hong és Page (2004) megmutatta, hogy a csoport teljesítményét nem elsősorban tagjainak egyéni képessége, hanem funkcionális sokfélesége határozza meg, vagyis az, hogy tagjai mennyire eltérően közelítenek a feladathoz és mennyire eltérő eljárásokkal keresik a megoldást. Fő eredményük szerint egy elegendően nagy jelöltállományból véletlenszerűen kiválasztott csoport felülmúlhatja az egyénileg legjobban teljesítők csoportját, mert a legjobbak egyre inkább hasonlítanak egymásra, és jobb képességeiket nem ellensúlyozza sokféleségük hiánya. Page (2007) ezt az összefüggést azonosság formájában is megfogalmazta: a kollektív négyzetes hiba az átlagos egyéni négyzetes hiba és a becslések szórásában mért sokféleség különbsége.

Vizsgálatom szempontjából a legfontosabb pontosítás Davis-Stober és szerzőtársai (2014) nevéhez fűződik, akik megmutatták, hogy a sokaság bölcsessége nem a függetlenségen, hanem az ítéletek különbözőségén múlik, és rendkívül robusztus még torzított, egymással korreláló becslések mellett is. Ennek jelentősége akkor válik világossá, amikor a sokaság tagjai nagy nyelvi modellek, amelyeknél a klasszikus értelemben vett függetlenség biztosan nem teljesül.

Az aggregálás módszerei

Ha a kimenet az aggregálás műveletén múlik, nem közömbös, hogyan összegezzük az egyéni ítéleteket. A klasszikus aggregátorok többsége az úgynevezett kvázi-aritmetikai átlagok családjába tartozik (Bullen, 2003), amelyek az egyéni becsléseket transzformációs függvénnyel képezik le, az új skálán átlagolnak, majd visszatranszformálják az átlagot. Attól függően, hogy milyen függvényt választunk, számtani, mértani, harmonikus vagy log-esély átlagot kapunk; ez utóbbi tulajdonságait a módszertani részben ismertetem. Galton eredeti választása miatt külön figyelmet érdemel a medián.

Az aggregátorok működési mechanizmusáról Satopää és szerzőtársai (2023) torzítás-információ-zaj (bias–information–noise) modellje ad számot. Geopolitikai előrejelzési tornák adatainak használatával megmutatták, hogy a különböző eljárások három eltérő úton javíthatják a pontosságot: a zaj csökkentésével, a torzítás csökkentésével, illetve a becslők között szétszórt információ kinyerésével. Két megállapításuk a jelen vizsgálat szempontjából is lényeges. Egyrészt az egyszerű átlagolás csaknem kizárólag a zaj csökkentésén keresztül hat, sőt a korlátos valószínűségi skálán a becsléseket a felezési pont felé húzva akár növelheti is a torzítást. Másrészt a log-esély térben végzett átlagolás ezt a torzítást hangolható paraméter nélkül korrigálja.

Nagy nyelvi modellek mint előrejelzők

A nyelvi modellek előrejelzési képességét vizsgáló kutatási irány kiindulópontja egy negatív eredmény. Amikor a GPT-4-et egy élő, háromhónapos Metaculus-tornán önállóan tesztelték, valószínűségi előrejelzései szignifikánsan pontatlanabbnak bizonyultak a humán sokaság mediánjánál, és a modell azt az egyszerű alapstratégiát sem tudta felülmúlni, amely minden kérdésre ötven százalékot tippel (Schoenegger & Park, 2023). Egyetlen modell tehát nem versenyképes az emberi sokasággal.

A helyzet két módon javítható. Az egyik a modellt kiegészítő eszközökkel ruházza fel (egy internetes keresésre és strukturált érvelésre épülő rendszerrel, amely a releváns információt aktívan felkutatja), ami így már megközelítette, sőt bizonyos esetekben felül is múlta az emberi sokaság pontosságát (Halawi és mtsai., 2024). A másik az aggregálás elvét alkalmazza a modellekre. Tizenkét különböző nyelvi modellből álló együttest építve, harmincegy valós Metaculus-kérdésen az együttes mediánja statisztikailag nem volt megkülönböztethető 925 emberi előrejelző mediánjától (Schoenegger és mtsai., 2024). A szerzők ugyanakkor kizárólag a mediánt alkalmazták aggregátorként, és a modelleket webes felületen keresztül kérdezték le, ami nem tette lehetővé a mintavételi paraméterek szabályozását.

A kutatási rés

A gépi sokaságokat vizsgáló empirikus munkák eddig nem tették fel a kérdést, hogy a klasszikus feltételek egyáltalán teljesülnek-e a nyelvi modellekből álló sokaságokban, és ha nem, az mennyiben korlátozza az elérhető nyereséget. Schoenegger és szerzőtársai (2024) például a gépi sokaság teljesítményét mérték az emberihez képest, de a sokaság belső szerkezetét nem vizsgálták.

Jelen tanulmány három kérdést tesz fel. Az első a tulajdonképpeni replikáció: reprodukálható-e a Galton által leírt aggregálási hatás egy olyan sokaságon, amelynek tagjai bizonyíthatóan nem függetlenek egymástól, és hogyan viszonyul a hatás mértéke az eredeti kísérletéhez. A második a sokaság belső szerkezetére irányul: mi adja a mesterségesen előállított sokaság sokféleségét, és az elérhető nyereség mennyire múlik a sokaság összetételén és méretén. A harmadik a hatás korlátaira kérdez rá: azonosítható-e olyan kérdéstípus, amelyen a gépi sokaság szisztematikusan alulteljesít, és ha igen, mi ennek a mechanizmusa.

Módszerek és adatok

A kérdések kiválasztása és az információszivárgás kizárása

Az elemzés alapját 47, a Metaculus előrejelzési platformon közzétett kérdés adja, amelyek 2026. július vége és augusztus vége között zárultak le. A Metaculus (é. n.) olyan nyilvános platform, ahol a regisztrált felhasználók valószínűségi becsléseket adnak jövőbeli eseményekre, a platform pedig ezekből egy súlyozott közösségi előrejelzést (community prediction) képez. A kérdések tematikailag vegyesek: makrogazdasági mutatók, geopolitikai események, technológiai és piaci fejlemények egyaránt szerepelnek köztük.

A kérdéskészlet összetétele típus szerint: 28 bináris (igen/nem kimenetelű), 9 numerikus, 5 diszkrét és 5 többválasztós kérdés. Az egyes kérdésekre becslést adó emberi előrejelzők száma erősen szórt: a medián 18 fő, a tartomány 2 és 160 fő között mozog. Az összes vizsgált kérdés megtalálható a Függelék F1. táblázatában.

A nagy nyelvi modellek előrejelzési képességét vizsgáló kutatásokban az információszivárgásnak két, egymástól független csatornája ismert. Az első a parametrikus szivárgás: ha a vizsgált kérdés a modell tanítási határideje előtt zárult le, a kimenetel bekerülhetett a tanítóadatba, így a mért teljesítmény nem előrejelzés, hanem felidézés. Douven (2026) kimutatta, hogy ennek mértéke nem elhanyagolható: a nagy és kis modellek közötti látszólagos képességbeli különbség 35,8 százalékról 8,9 százalékra csökkent, amint az elemzést kizárólag a modellek tanítási határidején túl lezáruló kérdésekre szűkítette.

A második a visszakeresési szivárgás, ami akkor lép fel, ha a modell webes keresésre is képes, és a kérdést utólag, a kimenetel ismertté válása után tesszük fel neki. A szakirodalom bevett védekezése erre a keresés időbeli szűrése, ezt azonban Paleka és szerzőtársai (2025) a nyelvi modellek előrejelzési képességét vizsgáló kutatások egyik visszatérő módszertani buktatójaként azonosítják. El Lahib és szerzőtársai (2026) megmutatták, hogy ez a védekezés megbízhatatlan. A Google before: szűrőjét 393 lezárt előrejelzési kérdésen auditálva a kérdések 71 százalékánál legalább egy visszakeresett oldal tartalmazott a határidő utáni lényegi információt, 41 százalékuknál pedig akadt olyan oldal, ami közvetlenül elárulta a választ; a DuckDuckGo esetében ezek az arányok 81, illetve 55 százalék voltak. A gyakorlati következmény jelentős: a szivárgó dokumentumokkal dolgozó modell Brier-pontszáma 0,10 volt, a szivárgásmentes dokumentumokkal viszont 0,24, vagyis a hiba több mint kétszeresére

nőtt. A szivárgás mechanizmusai hétköznapiak és nehezen kivédhetők: utólag frissített cikkek, kapcsolódó tartalmakat ajánló modulok, megbízhatatlan időbélyegek.

Jelen vizsgálat kialakítása a szivárgás kérdését azzal zárja ki, hogy a gépi becslést minden kérdésnél a lezárás napján, az esemény bekövetkezése előtt végzi el. Ekkor a kimenetel még nem létezik, így egyik szivárgási csatorna sem releváns. Ezt a kialakítást a fentebb ismertetett visszakeresési problémák indokolják. (Megjegyzem, hogy a lekérdezések egyébként sem tartalmaztak keresési bővítményt, de ennek a fentiek fényében csak másodlagos jelentősége van.)

Az emberi viszonyítási alapot a Metaculus közösségi előrejelzése adja. Ezt minden kérdésnél ugyanazon a napon rögzítettem, amikor a gépi előrejelzések készültek (17 esetben a kérdés lezárásakor kapott értéket használom, de a lezárás napja egybeesik a gépi becslés készítésének napjával). Így a két sokaság azonos információs horizontról dolgozik, és az összehasonlításnál a két sokaság közötti információs aszimmetria nem lehet hangsúlyos tényező.

A gépi sokaság előállítása

A gépi sokaság összeállításánál három elvet követtem, azzal a céllal, hogy a sokaságon belüli változatosság ne véletlenszerű zaj legyen, hanem módszertanilag kontrollált forrásokból származzon.

Az első választási elv a modellválasztás. Négy, eltérő fejlesztőtől származó és eltérő architektúrájú nagy nyelvi modell szerepelt a vizsgálatban: az OpenAI GPT-4o-mini, az Anthropic Claude Haiku 4.5, a Google Gemini 3 Flash és a Meta Llama 3.1 70B Instruct modellje. A négy modell elérése egységesen, egyetlen közvetítő felületen (OpenRouter) keresztül történt, így a lekérdezés formátuma minden modellnél azonos volt.

A második elv a szerepkör, vagyis a promptba illesztett személyleírás. Nyolc változat szerepelt: egy semleges kontroll (szerepmegjelölés nélkül), valamint hét szerep: közgazdász, jegybanki elemző, kötvénykereskedő, újságíró, kisvállalkozó, statisztikai elemző és szkeptikus laikus. A szerepkör-alapú becslés (elicitáció) szakirodalmi előzménye a szintetikus válaszadókkal (silicon samples) végzett kutatás: Argyle és szerzőtársai (2023) kimutatták, hogy a nyelvi modellek demográfiai jellemzőkkel való kondicionálása segít abban, hogy valós emberi alcsoportok válaszmegoszlásait reprodukáljuk. Ezt a jelenséget a szerzők algoritmikus hűségnek (algorithmic fidelity) nevezték el.

A harmadik elv a hőmérséklet (temperature), a nyelvi modellek mintavételi paramétere. A modell a következő szó, ill. token kiválasztásakor nem egyetlen választ számol ki, hanem valószínűségi eloszlást a lehetséges folytatásokról; a hőmérséklet ezt az eloszlást élesíti vagy lapítja a mintavétel előtt. Alacsony érték mellett a kimenet kiszámíthatóbb, magas érték mellett változatosabb. A hőmérséklet nem a modellnek szóló utasítás, hanem az alkalmazásprogramozási felületnek átadott, a prompttól elkülönülő paraméter. A vizsgálatban két beállítás szerepelt, 0,7 és 1,0. Nulla hőmérsékletet azért nem alkalmaztam, mert ott az ismétlések azonos választ adnának, és értelmüket vesztenék.

E három elven túl minden kombináció három ismétlésben szerepelt. Így a kialakítás kérdésenként $4 \times 8 \times 2 \times 3 = 192$ előrejelzést ír elő, ami a 47 kérdésre 9024 tervezett lekérdezést jelent. Ebből 112 kombináció (1,24 százalék) technikai okokból egyáltalán nem került rögzítésre, tehát 8912 válasz állt elő; ezek közül további 8 esetben nem született értelmezhető válasz. Az elemzés alapja tehát 8904 felhasználható gépi előrejelzés.

Az adatgyűjtés teljes egészében szkriptelt és automatizált volt. Minden egyes előrejelzésnél rögzítésre került a modell, a szerepkör, a hőmérséklet, az ismétlés sorszáma, a modell teljes szöveges válasza, a válaszból kinyert numerikus érték és az időbélyeg. A kimeneti formátum

rögzített volt (a válasz utolsó sorában PROBABILITY: X.XX alakban), így az értékek kinyerése gépi alapon történhetett, és nem volt szükség emberi beavatkozásra, értelmezésre.

Pontozás és aggregálás

A bináris kérdéseknél a pontosság mérőszáma a Brier-pontszám (Brier, 1950), vagyis az előrejelzett valószínűség és a tényleges kimenetel (0 vagy 1) különbségének négyzete. A mérőszám büntetés jellegű: az alacsonyabb érték jelent jobb teljesítményt, a nulla a tökéletes előrejelzést. A numerikus és diszkrét kérdéseknél a Brier-pontszám nem értelmezhető, ezért ezeknél az abszolút hibát használom, vagyis a becslés és a valós érték különbségének abszolút értékét. A többválasztós kérdéseket a formális pontozásból kihagytam: az elicitáció ezeknél csak a legvalószínűbbnek tartott válaszhoz rendelt egyetlen valószínűséget rögzítette, nem a teljes kategóriák fölötti eloszlást, így korrekt többosztályos pontszám nem számítható belőlük.

A gépi sokaság aggregált becslésének kiszámításához – Galton eredeti választásával összhangban – elsősorban a mediánt használom. Emellett összehasonlításképpen három további aggregátort is használok: a számtani átlagot, a 10 százalékos nyesett átlagot, valamint a (csak a bináris kérdéseknél értelmezhető) log-esély átlagot.

Ez utóbbi nem a valószínűségeket átlagolja közvetlenül, hanem azok esélyhányadosának logaritmusát: minden egyes becslést $p/(1-p)$ esélyre számít át, ennek logaritmusát átlagolja, majd az eredményt visszaszámítja valószínűséggé. A logit-transzformáció a 0 és 1 közötti intervallumot a teljes valós számegyenesre feszíti ki, így a szélsőértékekhez közeli becslések nagy súlyt kapnak. A szakirodalom ezt szélsőségesítésnek nevezi, és a valószínűségi aggregálás normatív mércéjeként tartja számon, feltéve, hogy az egyéni becslések feltételesen függetlenek és jól kalibráltak (Genest & Zidek, 1986; Morris, 1983).

A minta, korlátok és az adatbázis megosztott használata

Valamennyi kérdés hivatalos lezárása is elérhető a Metaculus oldalon, így az emberi és a gépi sokaság becslései nemcsak egymáshoz, hanem a valós kimenethez is hasonlíthatók. A pontozható részminta ennek ellenére szűkebb a teljes kérdéskészletnél. A 28 bináris kérdés mindegyike pontozható. A 14 numerikus és diszkrét kérdés közül 13 az; a kivételnél (az Egyesült Királyság Munkáspártjának heti felmérésátlaga) a Metaculus a kimenetelt a kérdés felső határa fölötti értékként rögzítette, ami ellehetetleníti a numerikus pontozást. Az 5 többválasztós kérdést a fent ismertetett okból nem pontozom, így végül 41 kérdés alapján bemutatom az eredményeket.

Két korlát azonosítható a mintában. Az első a méret: 28 bináris, illetve 13 numerikus kérdés kis minta, ezért az eredményeket feltáró jellegűként, nem megerősítő tesztként értelmezem. A vizsgálatot előzetesen nem regisztráltam, így minden itt közölt együttható és szignifikanciaszint leíró jellegű. A második, az értelmezés szempontjából fontosabb korlát a kimenetelek egyenetlensége: a 28 bináris kérdésből 23 nemleges és 5 igenlő kimenetelű lett. Ez azt jelenti, hogy egy olyan előrejelző, aki minden kérdésre egyszerűen nemmel válaszol, ezen a mintán jó Brier-pontszámot érne el. Az eredményeket ezért mindig a megfelelő viszonyítási alapok mellett kell értelmezni, és az alacsony igen-arány a szélsőséges kérdésekre vonatkozó megállapítások értékelésénél is szerepet kap.

Végül megjegyzendő, hogy az itt bemutatott adatgyűjtés egy másik, eltérő kérdésfelvetésű tanulmánynak is alapjául szolgál, amely a gépi sokaságot a szakértői panelek helyettesíthetőségének jövőkutatási kontextusában vizsgálja.

Elemzés és értékelés

Javít-e az aggregálás a bináris előrejelzéseken?

Az első kérdés az, hogy kimutatható-e egyáltalán aggregálási nyereség a gépi sokaságon. A mérés logikája Galtonét követi: minden kérdésnél összehasonlítom a sokaság mediánjának Brier-pontszámát azzal, amennyit a sokaság tagjai egyenként, átlagosan elértek. A vizsgálat alapja a bináris kérdések csoportja (B01–B28).

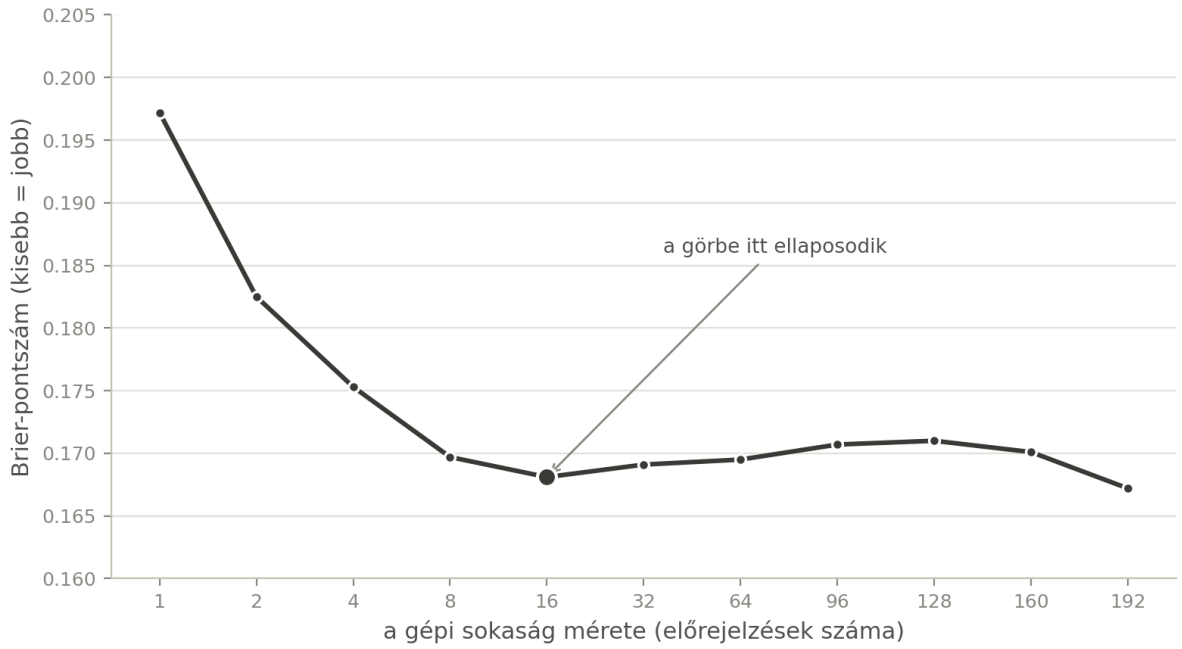
A 28 bináris kérdésnél az aggregátum átlagos Brier-pontszáma 0,178, míg az egyéni előrejelzések átlagos Brier-pontszáma 0,211. A különbség 0,033 Brier-pont az aggregátum javára, 95 százalékos bootstrap konfidenciaintervalluma 0,009 és 0,055 közé esik. A 28 kérdésből 22-nél bizonyult pontosabbnak az aggregátum, ami előjeltereszttel $p = 0{,}004$ szignifikanciaszintnek felel meg. A nyereség tehát statisztikailag megalapozott, de nagyságrendje óvatosságra int. A 0,033 Brier-pontos javulás mintegy 15 százalékos relatív hibacsökkentésnek felel meg. Ez kimutatható, de messze nem az a látványos hatás, amit Galton tapasztalt.

Az eredmény nem a medián választásán múlik. A másik három aggregátor is javít az egyéni átlagon, sőt átlagosan valamivel jobb Brier-pontszámot ér el a mediánnál: a számtani átlagé 0,172, a 10 százalékos nyesett átlagé 0,171, a log-esély átlagé 0,163. A mediánhoz mért eltérések azonban kicsik, és előjeltereszttel egyik sem szignifikáns ($p = 0{,}09$, 0,18, illetve 0,34). A log-esély átlag előnye szinte teljes egészében a szélsőséges kérdéseken jelentkezik (0,124, szemben a medián 0,150-es értékével). Ez összhangban áll Satopää és szerzőtársai (2023) eredményével, mely szerint a log-esély térben végzett átlagolás ellensúlyozza a felezési pont felé húzó torzítást, noha normatív feltétele, a becslők feltételes függetlensége, itt nyilvánvalóan nem teljesül.

Jelentős a kérdésenkénti szóródás mértéke. A legnagyobb nyereséget három olyan kérdésnél mértem, ahol a sokaság által adott becslések erősen szóródtak, de a medián a helyes irányba mutatott: a gázai stabilizációs erők telepítése (0,138), az albán kormány lemondása (0,125) és George Santos letartóztatása (0,115). A három legnagyobb veszteség ezzel szemben olyan kérdéseknél adódott, ahol a sokaság egységesen tévedett, így az aggregálás a téves konszenzust erősítette: az amerikai üzemanyagár-csúcs (–0,112), a Trump–Putyin telefonhívás (–0,080) és a júliusi hőrekord (–0,061). Ez a mintázat előrevetíti a gépi előrejelzés gyenge pontjairól szóló megállapítást: az aggregálás nem javítja a becslést, ha a tévedés az egész sokaságra, vagy legalábbis annak többségére kiterjed.

Érdemes megvizsgálni azt is, hány előrejelzésre van szükség az aggregálási nyereség eléréséhez. Az 1. ábra a sokaság méretének függvényében mutatja az aggregátum pontosságát. A görbe minden pontjához 300 alkalommal sorsoltam ki visszatevés nélkül a vízszintes tengelyen megadott számú előrejelzést, vettem a mediánjukat, és pontoztam a tényleges kimenetelhez képest; a számítás csak azokra a kérdésekre terjed ki, ahol mind a 192 előrejelzés rendelkezésre állt (17 kérdés), hogy a görbe minden pontja azonos kérdéskészleten alapuljon.

A Brier-pontszám monoton módon javul egészen 16 előrejelzésig, azon túl viszont a görbe ellaposodik: 16 és 192 között az értékek 0,167 és 0,171 között ingadoznak. A csökkenő hozadék klasszikus alakja tehát a gépi sokaságon is megjelenik, és a vizsgálatban alkalmazott 192-es sokaságméret jóval nagyobb volt annál, mint amit a hatás kimutatása igényelt.



1. ábra. A gépi sokaság mediánjának Brier-pontszáma a sokaság méretének függvényében (17 bináris kérdés, 300 újramintázás pontonként)

Forrás: saját szerkesztés

A tulajdonképpeni Galton-replikáció: numerikus becslések

Galton eredeti kísérlete folytonos mennyiség becslésére vonatkozott, nem valószínűségi előrejelzésre. Ebben az értelemben a numerikus és diszkrét kérdések állnak a legközelebb az 1906-os plymouth-i helyzethez, és az itt kapott eredményt tekintem a vizsgálat tulajdonképpeni replikációjának. Az elemzés a 13 pontozható kérdésen alapul (N01–N09, valamint D01, D02, D04 és D05); a D03 kódú kérdés a korábban ismertetett okból kimarad.

Az aggregálási nyereség itt is kimutatható: a 13 kérdésből 9-nél volt kisebb a gépi medián hibája az egyéni becslések átlagos hibájánál, és a javulás mediánja 13,8 százalék (1. táblázat). A szóródás nagy. A legerősebb javulás Donald Trump jóváhagyottsági mutatójánál jelentkezett (D02, 84,2 százalék), ezt követi a SpaceX augusztus végi árfolyama (N06, 76,1 százalék) és az Operation Epic Fury áldozatszáma (D04, 63,9 százalék). A négy kivétel mindegyike kicsi, a zaj nagyságrendjébe eső eltérés: az iráni háborúellenesség (D01, –2,3 százalék), az ausztrál munkanélküliségi ráta (N05, –5,2 százalék), a SpaceX júliusi árfolyama (N02, –5,5 százalék) és a réz határidős ára (N04, –9,5 százalék). Szisztematikus kudarcról tehát nem beszélhetünk, csak arról, hogy néhány kérdésnél az aggregálás nem hozott érdemi javulást.

1. táblázat. A numerikus és diszkrét kérdések eredményei kérdésenként

Kód	GÉPI: az egyéni becslések átlagos abszolút hibája	GÉPI: a sokaság mediánjának abszolút hibája	GÉPI: a sokaság számtani átlagának abszolút hibája	GÉPI: aggregálási nyereség (%) (medián vs. egyéni)	EMBERI: a közösségi előrejelzés abszolút hibája
D01	4,692	4,800	4,692	-2,3	1,300
D02	10,583	1,670	7,672	84,2	0,200
D04	27,693	10,000	27,672	63,9	

D05	8,120	7,000	8,120	13,8	
N01	0,305	0,196	0,183	36,0	0,006
N02	140,557	148,260	139,997	-5,5	0,260
N03	7,642	6,660	6,303	12,9	6,960
N04	0,153	0,168	0,152	-9,5	0,040
N05	0,161	0,169	0,147	-5,2	0,066
N06	20,150	4,810	3,285	76,1	8,690
N07	1,963	1,893	1,963	3,6	0,893
N08	64,503	32,169	47,268	50,1	21,831
N09	0,074	0,059	0,051	19,9	0,130

Megjegyzés: Az abszolút hibák az egyes kérdések saját mértékegységében értendők, ezért kérdések között közvetlenül nem összehasonlíthatók. Az aggregálási nyereség az egyéni becslések átlagos hibájához mért relatív javulás. A D03 kódú kérdés kimenetele a kérdés felső határa fölé esett, ezért numerikusan nem pontozható.

Forrás: saját szerkesztés

Megvizsgáltam azt is, hogy az egyéni becsléseknek hányad része került közelebb a valós értékhez, mint a sokaság mediánja. A 13 kérdés átlagában ez az arány 36,3 százalék, vagyis a medián az egyéni becslések csaknem kétharmadát (63,7 százalékát) múlta felül.

Galton közölt centilis-táblázatából ugyanez az arány kiszámítható (Galton, 1907). Az általa megadott medián 1207 font, a valós tisztított súly 1198 font volt, a medián hibája tehát 9 font. Egy egyéni tipp akkor pontosabb a mediánnál, ha 1189 és 1207 font közé esik; a közölt centilisek lineáris interpolációjával ez a sáv a 40,6. és az 50. centilis közé tehető. Galton 787 becslője közül tehát körülbelül 10 százalék, nagyjából 74 fő adott a mediánnál pontosabb tippet, vagyis a medián a sokaság mintegy 90 százalékát múlta felül. A gépi sokaságon ugyanez az arány 63,7 százalék. Hangsúlyozom, hogy a Galtonra vonatkozó érték interpolációval becsült közelítés, nem az eredeti adatokból számolt pontos szám; a két érték közötti különbség nagyságrendje azonban így is egyértelmű.

Végül érdemes kitérni arra a kérdésre, amit Galton kortársai is felvetettek. A közkeletű változat szerint a medián bizonyult a legjobb becslésnek. Wallis (2014) azonban Galton fennmaradt munkalapjait újraelemezve kimutatta, hogy a publikált adatokban átírási hiba szerepelt, és a javítás után a becslések számtani átlaga esett egybe pontosan a valós súllyal. Erre az anomáliára egyébként Hooker már közvetlenül a cikk megjelenése után felhívta a figyelmet egy olvasói levélben (Hooker, 1907).

A saját adataimon ez a kérdés sem dől el egyértelműen a medián javára. A 13 kérdésből mindössze 5-nél adott a medián pontosabb eredményt a számtani átlagnál, és a két aggregátor relatív eltérésének mediánja –5,7 százalék, vagyis tipikusan az átlag a valamivel pontosabb. A medián előnye máshol jelentkezik: amikor alulmarad, csak csekély mértékben (a legnagyobb hátrány 46 százalék), amikor viszont felülmúlja az átlagot, akkor nagy különbséggel teszi (78, 64 és 32 százalék). Ez pontosan az a robusztussági tulajdonság, ami miatt Galton a mediánt választotta: nem a tipikus esetben pontosabb, hanem a nagy tévedések ellen véd.

Az emberi sokasággal való összevetés ebben a kérdéscsoportban kevésbé kedvező a gépi sokaság számára. A 13 kérdésből 11-nél áll rendelkezésre emberi viszonyítási alap, és ezek közül mindössze 3-nál bizonyult pontosabbnak a gépi medián a közösségi előrejelzésnél. A kis elemszám miatt ebből messzemenő következtetést nem vonok le, de a szélsőséges kérdéseknél látott mintázattal összhangban áll.

Teljesülnek-e a sokaság klasszikus feltételei?

Ha az aggregálási nyereség a gépi sokaság esetén jóval szerényebb, mint Galtonnál, akkor kézenfekvő megvizsgálni, teljesülnek-e egyáltalán azok a feltételek, amelyeken a tömegek bölcsessége nyugszik. Ez az alfejezet két feltételt vizsgál: a sokféleséget és a függetlenséget. Az elemzés a bináris kérdések teljes körén (B01–B28) alapul.

Először azt vizsgáltam, honnan származik a gépi sokaság becsléseinek szóródása. A gépi sokaság előállításának három elve – modellválasztás, szerepkör, hőmérséklet – mindegyike szándékosan a változatosság növelését szolgálta, a hozzájárulásuk azonban rendkívül egyenetlen. A variancianalaízis alapján a modellválasztás magyarázza a szóródás 95 százalékát, a szerepkör 4,8 százalékot, a hőmérséklet pedig mindössze 0,2 százalékot (2. táblázat). A szerepkör-alapú és különösen a hőmérséklet-alapú változatosság alig járul hozzá a sokaság sokféleségéhez, tehát a gépi sokaságot a modellek különbözősége teszi elsősorban sokszínűvé.

2. táblázat. A gépi sokaság becsléseinek szóródása a három előállítási elv szerint

Az előállítás elve	Négyzetes eltérés (SS)	Szabadság-fok	A három elv szóródásának részesedése (%)	F	p
modellválasztás (4 modell)	62,91	3	95,0	794,38	<0,001
szerepkör (8 változat)	3,15	7	4,8	17,05	<0,001
hőmérséklet (2 beállítás)	0,15	1	0,2	5,78	0,016

Megjegyzés: Többszempontos varianciaanalízis (II. típusú négyzetösszegek) a bináris kérdések teljes körén (B01–B28), 5309 érvényes gépi előrejelzés alapján; a modell a kérdés azonosítóját is tartalmazza kontrollváltozóként. A részesedés a három elv egymáshoz viszonyított hozzájárulását mutatja, a kérdések közötti eltérés és a maradék szóródás nélkül.

Forrás: saját szerkesztés

A második feltétel, a függetlenség, még egyértelműbben sérül. A négy modell átlagos előjeles hibái között a páronkénti korrelációk 0,871 és 0,957 között mozognak, az átlaguk 0,90. A modellek tehát nem egymástól függetlenül tévednek, hanem jórészt ugyanazokon a kérdéseken és ugyanabba az irányba. Ez az az eredmény, ami a bináris kérdéseken mért szerény aggregálási nyereséget megmagyarázza: az átlagolás annyi hibát tud kioltani, amennyi a becslők között ténylegesen eltér, és a jelen sokaságban ez kevés.

A hat páronkénti érték közül öt szorosan egy sávban helyezkedik el (0,871 és 0,898 között), egy azonban kiugróan magas: a Llama 3.1 70B és a GPT-4o-mini hibái 0,957-es korrelációt mutatnak. Ez a két modell egyben a sokaság két leggyengébb tagja is: saját Brier-pontszámuk 0,286 és 0,242, szemben a Gemini 0,149-es és a Claude Haiku 0,157-es értékével. Két hasonlóan gyenge modell, amely ráadásul ugyanolyan típusú hibákat követ el, nem kioltja, hanem felerősíti egymás tévedését.

Ennek következménye közvetlenül mérhető. Ha a négy modellből egyet-egyet elhagyok, a teljes sokaság 0,178-es Brier-pontszáma a következőképpen változik: a Llama elhagyása 0,133-ra (–0,045), a GPT-4o-mini elhagyása 0,138-ra (–0,040) javítja az eredményt, míg a Claude Haiku elhagyása 0,213-ra (+0,034), a Gemini elhagyása 0,206-ra (+0,028) rontja. A sokaság tehát két olyan tagot tartalmaz, amelyek jelenléte összességében kárt okoz: egy kizárólag Claude Haikuból és Geminiből álló, feleakkora sokaság pontosabb előrejelzést adna, mint a teljes négymodelles együttes.

Ez az eredmény kapcsolódik a kiválasztott sokaság (select-crowd) irodalmához. Mannes és szerzőtársai (2014) megmutatták, hogy a becslők képesség szerinti előszűrése és csak a legjobbak átlagolása több környezetben robusztusabb, mint a teljes sokaság átlagolása vagy az

egyetlen legjobb becslőre való hagyatkozás. A jelen adatok ezzel összhangban állnak, két megszorítással: a szűrés itt utólagos, a kimenetel ismeretében történt, tehát nem előrejelzési eljárás, és négy modell alapján nem lehet általános szabályt megfogalmazni arról, hány és milyen modellt érdemes bevonni.

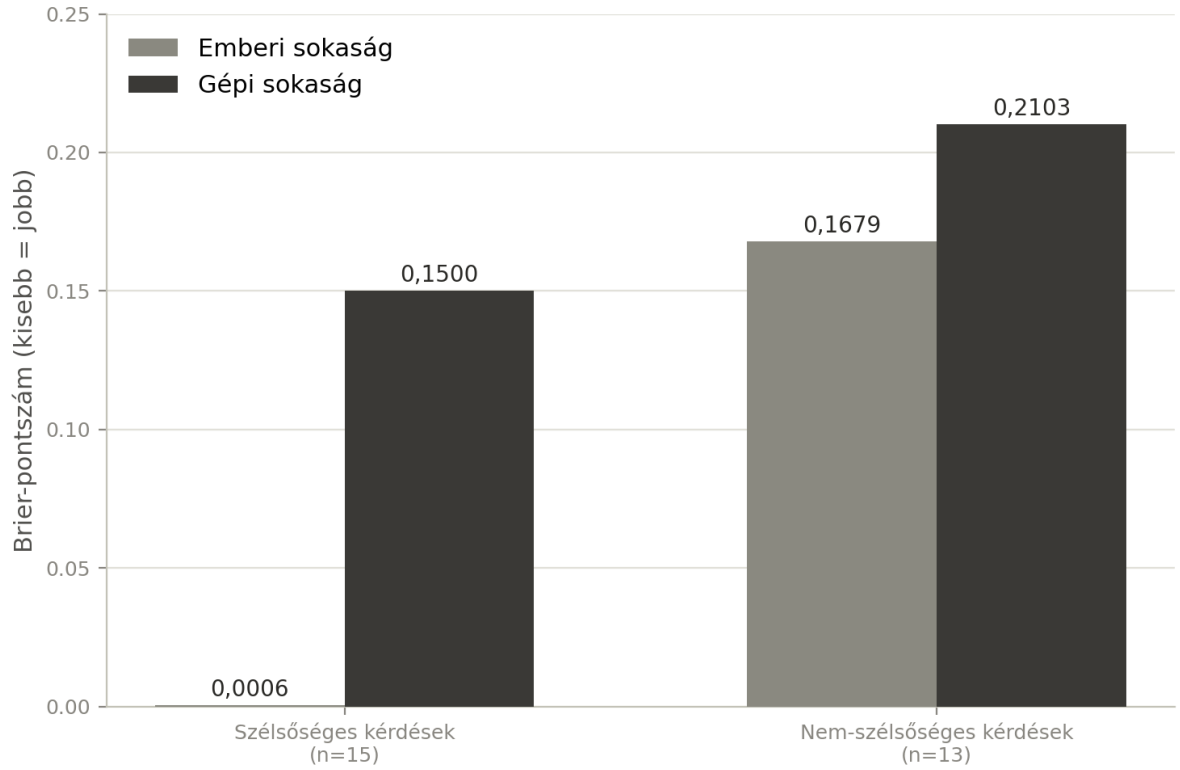
Ez nem mond ellent Davis-Stober és szerzőtársai (2014) eredményének: a bináris kérdéseken mért, statisztikailag megalapozott nyereség azt mutatja, hogy korreláló becslők mellett is van mit aggregálni. A korreláció mértéke azonban korlátozza, mennyi hibakioltás érhető el.

Hol romlik el a gépi előrejelzés?

A gépi és az emberi sokaság összehasonlítása a bináris kérdések teljes körén (B01–B28) egyértelmű képet mutat: a gépi sokaság átlagos Brier-pontszáma 0,178, az emberi közösségi előrejelzésé 0,078, a különbség 0,100 az emberi sokaság javára. A 28 kérdésből 24-nél bizonyult pontosabbnak az emberi becslés, ami előjeleszttel $p < 0{,}001$ szignifikanciaszintnek felel meg. Az emberi sokaság tehát összességében határozottan pontosabb.

Ez az összesített kép azonban elfedi, hogy a hátrány rendkívül egyenetlenül oszlik el. A kérdéseket két csoportra bontottam aszerint, mennyire volt bizonyos az emberi sokaság a lezárás pillanatában. Szélsőségesnek nevezem azokat a kérdéseket, ahol a közösségi előrejelzés 5 százalék alatt vagy 95 százalék fölött volt. 15 ilyen kérdés van az adatbázisban (B08–B11, B14, B17–B20, B22, B24–B28), a maradék 13 viszont valóban bizonytalan kimenetelű (B01–B07, B12, B13, B15, B16, B21, B23).

Amennyiben a fenti összehasonlítást a két csoporton külön-külön elvégezzük, markáns különbséget kapunk (2. ábra). A szélsőséges kérdéseken az emberi sokaság Brier-pontszáma 0,0006, tehát gyakorlatilag hibátlan, a gépié viszont 0,150. Mind a 14 értékelhető kérdésnél az emberi becslés a pontosabb ($p < 0{,}001$). A nem-szélsőséges kérdéseken a gépi és emberi becslés közötti rés 0,042, alig több mint negyede a szélsőséges kérdéseken mért 0,149-nek, és a különbség a szokásos szinten nem szignifikáns ($p = 0{,}092$).



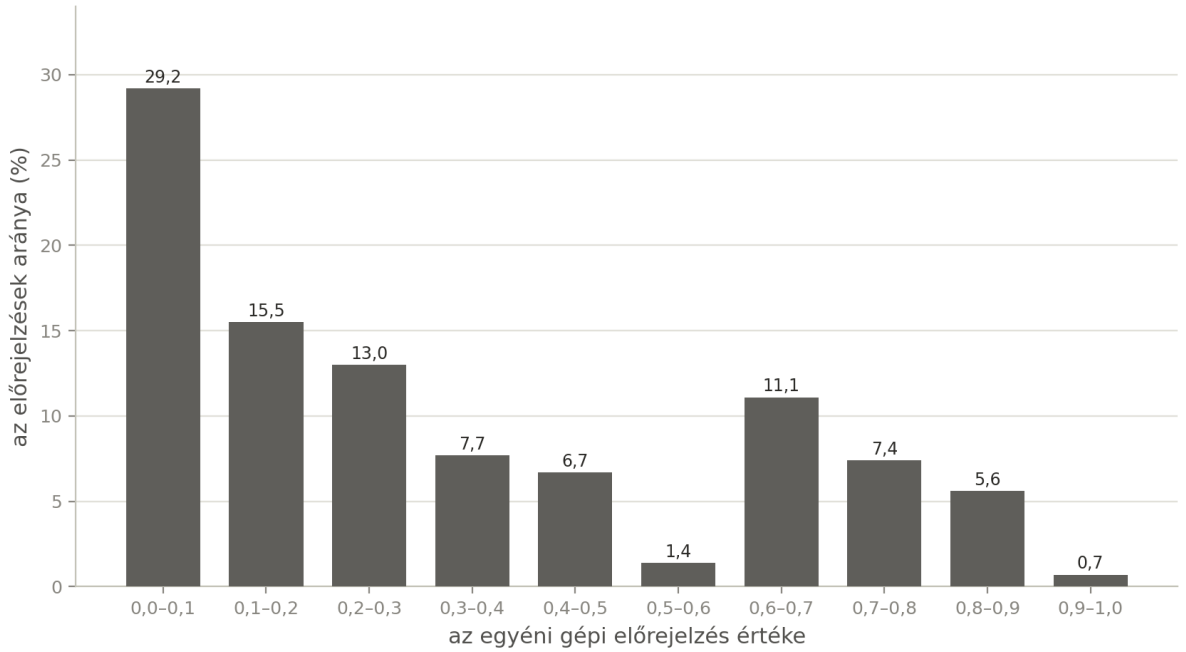
2. ábra. A gépi és az emberi sokaság Brier-pontszáma szélsőséges és nem-szélsőséges kérdéseken

Forrás: saját szerkesztés

A nem-szélsőséges csoport eredményét egyetlen kérdés uralja, ezért érdemes külön kiemelni. A B12 kódú kérdésnél (a búvárúszó-verseny leggyorsabb ideje) a gépi sokaság mediánja 0,80 volt, az emberi közösségi előrejelzés 0,318, a kimenetel pedig nemleges lett; az így adódó 0,539-es rés a 13 kérdés teljes réskülönbségének 98 százalékát adja. Ezt a kérdést elhagyva a nem-szélsőséges csoportban a rés 0,001-re, vagyis gyakorlatilag nullára csökken. A 13 kérdésre számított résmedián 0,050, ami szintén azt jelzi, hogy a csoportátlagot néhány szélsőséges eset húzza fel.

Óvatosan fogalmazva tehát megállapítható, hogy a bizonytalan kimenetelű kérdéseken a gépi sokaság nem bizonyult egyértelműen rosszabbnak az emberinél, a szélsőséges kérdéseken viszont nagyságrendi a hátránya. A gépi sokaság teljes látszólagos lemaradása lényegében ebben az egy kérdéscsoportban koncentrálódik.

Nem arról van szó, hogy a mesterséges intelligencia óvatosabb a szélsőséges kérdésekben. Ha a modellek egyszerűen bizonytalanok lennének, becsléseik 0,5 körül csoportosulnának. A 3. ábra ehelyett kétcsúcsú eloszlást mutat: a szélsőségesen alacsony valószínűségű kérdéseken adott egyéni gépi előrejelzések 29,2 százaléka helyesen a 0 és 0,1 közötti sávba esik, további 15,5 százalék a 0,1 és 0,2 közötti sávba, a 0,5 körüli tartomány viszont kifejezetten ritkán fordul elő (1,4 százalék). Ezzel szemben az előrejelzések 18,5 százaléka a 0,6 és 0,8 közötti sávba esik. Ezek nem bizonytalan, hanem magabiztosan téves becslések: a modellek egy jól elkülönülő része határozottan az ellenkező irányba tippelt, mint amerre a kimenetel alakult, ezért a válaszok eloszlása bimodális.



3. ábra. Az egyéni gépi előrejelzések eloszlása az 5 százalék alatti közösségi valószínűségű kérdéseken

Forrás: saját szerkesztés

Ha a szélsőséges kérdéseken mért hátrány a sokaság összetételéből fakad, akkor a gyenge tagok elhagyásának javítania kell az eredményt. Ezt közvetlenül megvizsgáltam; a korábban azonosított két gyenge modellt (Llama 3.1 70B és GPT-4o-mini) elhagyva a maradék kétmodelles sokaság a 15 szélsőséges kérdésen 0,150 helyett 0,062-es Brier-pontszámot ér el. A javulás mind a 15 kérdésnél jelentkezik ($p < 0,001$), és az emberi sokasághoz mért rést 0,149-ről 0,061-re csökkenti.

A különbség oka jól látható az egyéni becslések szintjén. A szélsőségesen alacsony valószínűségű kérdéseken a Claude Haiku átlagos becslése 0,131, a Geminié 0,244, a Llamáé viszont 0,477, a GPT-4o-minié 0,416, vagyis az utóbbi kettő átlagosan a bizonytalansági tartomány közepe körül tippel olyan kérdéseknél, ahol a helyes válasz a nullához közelít. Még szembetűnőbb a határozott elköteleződés hiánya: míg a Claude Haiku becsléseinek 33,6, a Geminiének 28,2 százaléka esik 0,05 alá, a Llamánál ez az arány 1,0, a GPT-4o-mininél pedig 0,1 százalék. A két gyengébb modell tehát gyakorlatilag képtelen a határozottan alacsony valószínűség kimondására, és ezzel húzza fel az egész sokaság becslését.

A modellek közötti különbségek összefüggenek a modellek általános képességszintjével. A 3. táblázat a négy modell Artificial Analysis Intelligence Index szerinti pontszámát veti össze saját előrejelzési pontosságukkal és a sokasághoz való hozzájárulásukkal (Artificial Analysis, é. n.). A két magasabb indexű modell (Gemini 28, Claude Haiku 24) egyben a két pontosabb előrejelző is. Négy adatponton ez nem tesztelhető statisztikailag, ezért szemléltető mintázatként közlöm.

3. táblázat. A négy nyelvi modell általános képességszintje és előrejelzési pontossága

Modell	Artificial Analysis Intelligence Index	Saját Brier-pontszám	A sokaság Brier-pontszáma a modell elhagyásával	Változás a teljes sokasághoz képest
Gemini 3 Flash (Google)	28	0,1492	0,2063	0,0283
Claude Haiku 4.5 (Anthropic)	24	0,1567	0,2125	0,0344
GPT-4o-mini (OpenAI)	7	0,2420	0,1382	-0,0398
Llama 3.1 70B (Meta)	7	0,2857	0,1328	-0,0453

Forrás: saját szerkesztés; a modellek képességindexe az Artificial Analysis (é. n.) adatai alapján (2026. szeptember 4-i lekérdezés)

Egy további érdekesség. Az egyéni pontosság és a sokasághoz való hozzájárulás nem ugyanaz: a Gemini saját Brier-pontszáma valamivel jobb a Claude Haikuénál, a sokaságból való elhagyása mégis kisebb kárt okoz (+0,028, szemben +0,034-gyel). A sokaság szempontjából tehát nem egyszerűen a legpontosabb tag a legértékesebb.

Simpson-paradoxon

A bináris kérdések teljes körén (B01–B28) egy további érdekes jelenség figyelhető meg, ami a gépi válasz hosszához, és az ún. Simpson-paradoxonhoz (Blyth, 1972; Simpson, 1951) kapcsolódik. A modellek minden előrejelzéshez rövid szöveges indoklást is adtak. A válasz hossza és az előrejelzés pontossága közötti kapcsolat előjelet vált attól függően, hogy kontrollálok-e a modellre. Az összevont adatokon a hosszabb érvelés jobb pontossággal jár együtt (Spearman-féle rangkorreláció: $r = –0,161$, $p < 0,001$, $n = 5309$). Modellen belül vizsgálva azonban mind a négy modellnél fordított a kapcsolat: a hosszabb válasz rosszabb előrejelzéssel társul ($r = +0,028$ és $+0,231$ között; a Claude Haiku kivételével mindegyik szignifikáns).

Az összevont eredmény tehát műtermék. A Claude Haiku a legbőbeszédűbb modell (átlagosan 1793 karakter) és egyben az egyik legpontosabb, a GPT-4o-mini és a Llama pedig rövidebben ír (1101, illetve 889 karakter) és pontatlanabb; az összevonás így a modellazonosságot mossa össze a válaszhosszal. A jelenség a Simpson-paradoxon tankönyvi esete, és jó példa arra, miért nem szabad a modelleket egyetlen közös sokaságként kezelni.

A modellen belüli összefüggésnek ugyanakkor van értelmezhető mechanizmusa. A hosszabb válaszok szisztematikusan kevésbé szélsőségesek: a válaszhossz és a 0,5-től mért távolság között mind a négy modellnél negatív a korreláció ($r = –0,056$ és $–0,289$ között). Amikor tehát egy modell többet ír, akkor a bizonytalansági tartomány közepe felé húz, ami ezen a mintán Brier-pontokba kerül.

Konklúzió

A tanulmány Galton 1906-os megfigyelését ismételte meg egy olyan sokaságon, amelynek tagjai nem emberek, hanem nagy nyelvi modellek. Négy modell, nyolc szerepkör és két hőmérséklet-beállítás kombinálásával kérdésenként 192 előrejelzést gyűjtöttem 47 valós, a lekérdezés pillanatában még le nem zárult Metaculus-kérdésre.

Az aggregálási hatás kimutatható, de lényegesen gyengébb, mint az eredeti kísérletben. A bináris kérdéseken (B01–B28) a sokaság mediánjának Brier-pontszáma 0,178, szemben az egyéni előrejelzések 0,211-es átlagával; a 0,033-es különbség 95 százalékos konfidenciaintervalluma 0,009 és 0,055 közé esik, és a 28 kérdésből 22-nél javított az aggregálás ($p = 0,004$). A relatív hibacsökkenés azonban mindössze 15 százalék. A numerikus és diszkrét kérdéseken (N01–N09, D01, D02, D04, D05) a kép hasonló: a 13 kérdésből 9-nél javított az aggregálás, a javulás mediánja 13,8 százalék. A Galtonnal közvetlenül összevethető mutató szerint a gépi sokaság mediánja az egyéni becslések 63,7 százalékát múlta felül, miközben Galton eredeti, 1906-os sokaságán ez az arány mintegy 90 százalék volt.

A különbséget a sokaság szerkezete magyarázza. A mesterségesen előállított sokféleség csaknem teljes egészében egyetlen forrásból származik: a modellválasztás a szóródás 95,6 százalékát adja, a szerepkör 4,2, a hőmérséklet pedig 0,2 százalékát. A függetlenség feltétele pedig egyértelműen sérül: a négy modell hibái között a páronkénti korrelációk 0,871 és 0,957 között mozognak, átlaguk 0,90. A sokaság tagjai tehát jórészt ugyanazokon a kérdéseken és ugyanabba az irányba tévednek, és az aggregálás csak annyi hibát tud kioltani, amennyi közöttük ténylegesen eltér.

Ebből következik a tanulmány gyakorlati szempontból leghasznosabb megállapítása: a gépi sokaságnál az összetétel fontosabb, mint a méret (a kérdésekre adott becslések száma). A sokaságméret-görbe már 16 előrejelzésnél ellaposodik, a 192-es méret tehát jóval több volt a szükségesnél. Ezzel szemben a tagok kiválasztása érdemben számít: a két gyengébb modell elhagyása 0,178-ról 0,133-ra, illetve 0,138-ra javítja a sokaság pontosságát, míg a két erősebbé 0,213-ra és 0,206-ra rontja. Egy feleakkora, de gondosan összeállított sokaság pontosabb, mint a teljes együttes. Az eredmények arra utalnak, hogy azok a modellek adnak jobb becslést, amelyek az általános képességszintet mérő teszteken is jobban teljesítenek; négy modell alapján ez azonban csak szemléltető mintázat.

A gépi sokaság egyértelműen elmarad az emberi közösségi előrejelzéstől: a 28 kérdésből 24-nél az emberi becslés volt a pontosabb ($p < 0,001$). A hátrány azonban rendkívül egyenetlenül oszlik el. A szélsőséges, csaknem bizonyos kimenetelű kérdéseken az emberi sokaság Brier-pontszáma 0,0006, a gépié 0,150, és mind a 14 értékelhető kérdésnél az emberi becslés a pontosabb. A bizonytalan kimenetelű kérdéseken viszont a rés 0,042-re szűkül, és a szokásos

szinten nem szignifikáns (p = 0,092). A gépi sokaság teljes látszólagos lemaradása lényegében egyetlen kérdéstípusban koncentrálódik.

A szélsőségesen alacsony valószínűségű kérdéseknél az egyéni gépi becslések eloszlása kétcsúcsú: 44,7 százalékuk helyesen a 0,2 alatti tartományba esik, a 0,5 körüli sáv viszont csaknem üres (1,4 százalék), miközben 18,5 százalékuk a 0,6 és 0,8 közötti tartományban, tehát magabiztosan az ellenkező irányba tippel. Ez a mintázat is a sokaság összetételére vezethető vissza: a két gyenge modell elhagyásával a szélsőséges kérdéseken mért Brier-pontszám 0,150-ről 0,062-re javul, mind a 15 kérdésnél (p < 0,001), és az emberi sokasághoz mért rés a felére-harmadára csökken.

Néhány megjegyzés a tanulmány korlátairól. A minta kicsi, a kimenetelek erősen egyenetlenek, és négy modell alapján a modellválasztásra vonatkozó megállapítások nem általánosíthatók. Három irány kínálkozik a folytatásra. Az első a kiegyensúlyozottabb kérdéskészlet: a jelen vizsgálat kötő korlátja nem a mintaméret volt, hanem az igenlő kimenetelek hiánya. A második a sokaság összetételének előzetes, a kimenetel ismerete nélküli optimalizálása, a kiválasztott sokaság irodalmának szellemében. A harmadik az időhorizont változtatása: itt minden becslés a lezárás napján készült, és nyitott kérdés, hogy hosszabb horizonton a gépi sokaság gyorsabban vagy lassabban veszít-e pontosságából, mint az emberi.

Irodalomjegyzék

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337–351. https://doi.org/10.1017/pan.2023.2

Artificial Analysis. (é. n.). *Intelligence Index v4.1.1*. Elérés: 2026. szeptember 4., https://artificialanalysis.ai

Blyth, C. R. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, *67*(338), 364–366. https://doi.org/10.1080/01621459.1972.10482387

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078%3C0001:VOFEIT%3E2.0.CO;2

Bullen, P. S. (2003). *Handbook of means and their inequalities*. Kluwer Academic Publishers.

Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When is a crowd wise? *Decision*, *1*(2), 79–101. https://doi.org/10.1037/dec0000004

Douven, I. (2026). *Wisdom of LLM crowds: Aggregation and contamination in language model ensembles* (arXiv:2607.18269). arXiv. https://doi.org/10.48550/arXiv.2607.18269

El Lahib, A., Xia, Y.-J., Li, Z., Wang, Y., & Pi, X. (2026). *Temporal leakage in search-engine date-filtered web retrieval: A retrospective forecasting case study* (arXiv:2602.00758). arXiv. https://doi.org/10.48550/arXiv.2602.00758

Galton, F. (1907). Vox populi. *Nature*, *75*(1949), 450–451. https://doi.org/10.1038/075450a0

Genest, C., & Zidek, J. V. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, *1*(1), 114–135. https://doi.org/10.1214/ss/1177013825

Halawi, D., Zhang, F., Chen, Y.-H., & Steinhardt, J. (2024). *Approaching human-level forecasting with language models* (arXiv:2402.18563). arXiv. https://doi.org/10.48550/arXiv.2402.18563

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, *101*(46), 16385–16389. https://doi.org/10.1073/pnas.0403723101

Hooker, R. H. (1907). Mean or median. *Nature*, *75*, 487–488. https://doi.org/10.1038/075487a0
Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science*, *52*(1), 111–127. https://doi.org/10.1287/mnsc.1050.0459
Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology*, *107*(2), 276–299. https://doi.org/10.1037/a0036677
Metaculus. (é. n.). *Metaculus*. https://www.metaculus.com
Morris, P. A. (1983). An axiomatic approach to expert resolution. *Management Science*, *29*(1), 24–32. https://doi.org/10.1287/mnsc.29.1.24
Page, S. E. (2007). *The difference: How the power of diversity creates better groups, firms, schools, and societies*. Princeton University Press.
Paleka, D., Goel, S., Geiping, J., & Tramèr, F. (2025). *Pitfalls in evaluating language model forecasters* (arXiv:2506.00723). arXiv. https://doi.org/10.48550/arXiv.2506.00723
Satopää, V. A., Salikhov, M., Tetlock, P. E., & Mellers, B. (2023). Decomposing the effects of crowd-wisdom aggregators: The bias–information–noise (BIN) model. *International Journal of Forecasting*, *39*(1), 470–485. https://doi.org/10.1016/j.ijforecast.2021.12.010
Schoenegger, P., & Park, P. S. (2023). *Large language model prediction capabilities: Evidence from a real-world forecasting tournament* (arXiv:2310.13014). arXiv. https://doi.org/10.48550/arXiv.2310.13014
Schoenegger, P., Tuminauskaite, I., Park, P. S., Bastos, R. V. S., & Tetlock, P. E. (2024). Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy. *Science Advances*, *10*(45), eadp1528. https://doi.org/10.1126/sciadv.adp1528
Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *13*(2), 238–241. https://doi.org/10.1111/j.2517-6161.1951.tb00088.x
Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations* (1. kiad.). Doubleday.
Wallis, K. F. (2014). Revisiting Francis Galton's forecasting competition. *Statistical Science*, *29*(3), 420–424. https://doi.org/10.1214/14-STS468

Függelék

F1. táblázat. A vizsgálatban szereplő Metaculus-kérdések

Kód	Metaculus ID	Típus	A kérdés szövege	Lezárás	Előrejelzők
B01	43992	bináris	Will July 2026 be the warmest July on record, according to Copernicus?	2026-07-31	25
B02	44363	bináris	Will George Soros be the top Democratic-aligned donor on WaPo's leaderboard with at least 3x the #2 donor as of August 1, 2026?	2026-07-31	7
B03	44413	bináris	Will the FAO Food Price Index reach a new yearly high for July 2026?	2026-07-31	6
B04	43865	bináris	Will the U.S. Senate pass an Iran war powers resolution before August 1, 2026?	2026-07-31	23
B05	43822	bináris	Will Iceland vote Yes in the 2026 EU negotiations referendum?	2026-08-01	105

B06	43491	bináris	Will a new High Representative for Bosnia and Herzegovina officially be in place before September 2, 2026?	2026-08-01	149
B07	43844	bináris	Will Hakainde Hichilema receive at least 65% of the vote in the first round of Zambia's 2026 presidential election?	2026-08-12	12
B08	44792	bináris	Will transit calls in the Bab el-Mandeb Strait be cut by half or more on August 31, 2026 compared to July 2026?	2026-08-24	13
B09	44975	bináris	Will any personnel of the International Stabilization Force be deployed into Gaza before August 25, 2026?	2026-08-24	17
B10	44241	bináris	Will the Lincoln Memorial Reflecting Pool be fully repaired before September 2026?	2026-08-24	128
B11	44512	bináris	Will the Government of Albania or Prime Minister Edi Rama resign before September 2026?	2026-08-24	94
B12	44548	bináris	Will the fastest official time at the 2026 World Bog Snorkelling Championships be under 1:15.00?	2026-08-30	8
B13	44693	bináris	Will the desktop market share of Linux exceed 8% in August 2026?	2026-08-30	22
B14	44799	bináris	Will a bot finish in the top 5 of the Summer 2026 Metaculus Cup?	2026-08-30	23
B15	44973	bináris	Will August 2026 be the warmest August on record, according to Copernicus?	2026-08-30	23
B16	44239	bináris	Will Elon Musk be a trillionaire on August 31, 2026?	2026-08-30	24
B17	39886	bináris	Will Russia, India, and China hold a trilateral (three-way) meeting before September 1, 2026?	2026-08-31	25
B18	44798	bináris	Will Death Valley hit 125 degrees in August 2026?	2026-08-31	22
B19	44513	bináris	Will Donald Trump and Vladimir Putin speak on the phone in August 2026?	2026-08-31	40
B20	44230	bináris	Will the IAEA conduct an on-site inspection at any of Iran's Natanz, Fordow, or Isfahan nuclear sites before September 1, 2026?	2026-08-31	14
B21	44244	bináris	Will another company get dropped from the Dow Jones Industrial Average in July or August 2026?	2026-08-31	9
B22	44801	bináris	Will gas prices in the U.S. reach a new all-time high before September 6, 2026?	2026-08-31	19
B23	43437	bináris	Will there be a successful coup in Africa or Latin America before September 1, 2026?	2026-08-31	55
B24	43703	bináris	Will a hurricane make landfall in the continental United States before September 1, 2026?	2026-08-31	18
B25	43708	bináris	Will Starlink's website show South Africa as a country with Starlink availability before September 1, 2026?	2026-08-31	17
B26	43709	bináris	Will Starlink's website show India as a country with Starlink availability before September 2026?	2026-08-31	11
B27	43853	bináris	Will former U.S. Representative George Santos be arrested before September 1, 2026?	2026-08-31	14

B28	43985	bináris	Before September 1, 2026, will Turkish President Recep Tayyip Erdoğan be legally permitted to run for another term?	2026-09-01	24
D01	44694	diszkrét	What percent of Americans will oppose the Iran War on August 14, 2026?	2026-08-13	12
D02	44674	diszkrét	What will be Donald Trump's approval rating on August 14, 2026?	2026-08-13	12
D03	44672	diszkrét	What will UK Labour Party's 7-poll moving average percent be on August 31, 2026?	2026-08-24	89
D04	44802	diszkrét	How many deaths from Operation Epic Fury will the U.S. government officially report on August 31, 2026?	2026-08-30	2
D05	44451	diszkrét	What will be the approval rating of Vladimir Putin in August 2026?	2026-08-31	6
T01	44362	többválasztós	Who will be the largest Republican donor in the 2026 cycle, as of August 1, 2026?	2026-07-31	5
T02	44671	többválasztós	How many trillion dollar companies will there be on August 31, 2026?	2026-08-25	97
T03	43898	többválasztós	What will happen in the Russia–Ukraine war between July 20 and September 1, 2026?	2026-08-25	133
T04	43705	többválasztós	What will be the average 30-year fixed rate mortgage for the week ending August 27, 2026?	2026-08-26	8
T05	43842	többválasztós	What percentage of votes will be for Yes in Iceland's August 2026 EU referendum?	2026-08-28	8
N01	44165	numerikus	How many foreign visitors will fly to the United States in July 2026?	2026-07-31	7
N02	44168	numerikus	What will be the highest stock price of SpaceX in July 2026?	2026-07-31	10
N03	44551	numerikus	What percentage of the vote will Nigel Farage receive in the Clacton parliamentary by-election in August 2026?	2026-08-12	98
N04	43483	numerikus	What will be the highest price of copper futures in August 2026?	2026-08-14	98
N05	44932	numerikus	What will the seasonally adjusted unemployment rate be in Australia for July 2026?	2026-08-19	7
N06	43939	numerikus	What will SpaceX's stock price be at the end of August 2026?	2026-08-21	160
N07	44746	numerikus	What will be the seasonally adjusted S&P/Case-Shiller US National Home Price Index for June 2026?	2026-08-24	9
N08	44791	numerikus	What will be the weekly cumulative burnt areas (ha) from forest fires in the EU on August 26, 2026?	2026-08-25	80
N09	44717	numerikus	What will the seasonally adjusted U-3 unemployment rate be in the United States in August 2026?	2026-08-31	8

Forrás: saját szerkesztés a Metaculus (é. n.) adatai alapján