



Közzététel: 2025. március 28.

A tanulmány címe:

**Nagy nyelvmodellek által generált adatok társadalomtudományi használatának kritikai elemzése**

Szerző:

**RAKOVICS ZSÓFIA**

az Eötvös Loránd Tudományegyetem (ELTE) Szociológia Doktori Iskolájának doktoranduszhallgatója és az ELTE Társadalomtudományi Kar Társadalomkutatások Módszertana Tanszék egyetemi tanársegédje és kutatója

E-mail: [zsafia.rakovics@tatk.elte.hu](mailto:zsafia.rakovics@tatk.elte.hu)

DOI: <https://doi.org/10.20311/stat2025.03.hu0259>

**Az alábbi feltételek érvényesek minden, a Központi Statisztikai Hivatal (a továbbiakban: KSH) Statisztikai Szemle c. folyóiratában (a továbbiakban: Folyóirat) megjelenő tanulmányra. Felhasználó a tanulmány vagy annak részeit felhasználásával egyidejűleg tudomásul veszi a jelen dokumentumban foglalt felhasználási feltételeket, és azokat magára nézve kötelezőnek fogadja el. Tudomásul veszi, hogy a jelen feltételek megszegéséből eredő valamennyi kárért felelősséggel tartozik.**

1. A jogszabályi tartalom kivételével a tanulmányok a szerzői jogról szóló 1999. évi LXXVI. törvény (Szjt.) szerint szerzői műnek minősülnek. A szerzői jog jogosultja a KSH.
2. A KSH földrajzi és időbeli korlátozás nélküli, nem kizárólagos, nem átadható, térítésmentes felhasználási jogot biztosít a Felhasználó részére a tanulmány vonatkozásában.
3. A felhasználási jog keretében a Felhasználó jogosult a tanulmány:
  - a) oktatási és kutatási célú felhasználására (nyilvánosságra hozatalára és továbbítására a 4. pontban foglalt kivétellel) a Folyóirat és a szerző(k) feltüntetésével;
  - b) tartalmáról összefoglaló készítésére az írott és az elektronikus médiában a Folyóirat és a szerző(k) feltüntetésével;
  - c) részletének idézésére – az átvevő mű jellege és célja által indokolt terjedelemben és az eredetihez híven – a forrás, valamint az ott megjelölt szerző(k) megnevezésével.
4. A Felhasználó nem jogosult a tanulmány továbbértékesítésére, haszonszerzési célú felhasználására. Ez a korlátozás nem érinti a tanulmány felhasználásával előállított, de az Szjt. szerint önálló szerzői műnek minősülő mű ilyen célú felhasználását.
5. A tanulmány átdolgozása, újra publikálása tilos.
6. A 3. a)–c) pontban foglaltak alapján a

Folyóiratot és a szerző(ke)t az alábbiak szerint kell feltüntetni:

„*Forrás: Statisztikai Szemle* c. folyóirat 103. évfolyam 3. számában megjelent, **Rakovics Zsófia** által írt, **Nagy nyelvmodellek által generált adatok társadalomtudományi használatának kritikai elemzése** című tanulmány (link csatolása)”

7. A Folyóiratban megjelenő tanulmányok kutatói véleményeket tükröznek, amelyek nem feltétlenül esnek egybe a KSH vagy a szerzők által képviselt intézmények hivatalos álláspontjával.

Rakovics Zsófia

# Nagy nyelvmodellek által generált adatok társadalomtudományi használatának kritikai elemzése\*

## Critical analysis of the use of data generated by large language models in the social sciences

Rakovics Zsófia, az Eötvös Loránd Tudományegyetem (ELTE) Szociológia Doktori Iskolájának doktorandusz-hallgatója és az ELTE Társadalomtudományi Kar Társadalomkutatások Módszertana Tanszék egyetemi tanársegédje és kutatója  
E-mail: zsofia.rakovics@tatk.elte.hu

A nagy nyelvmodellek (*Large Language Models*, LLM) megjelenése új lehetőséget teremt a társadalomkutatás módszerei szempontjából. Minden olyan kutatás esetében, ahol a nyelv közvetíti az empiriát, reális lehetőséggé válik valós személyek helyett vagy mellett LLM-ek által szimulált virtuális válaszadók bevonása adatközlőként. A mesterséges intelligencia által generált adatokban rejlő potenciál olyan mértékű, hogy annak módszertani és kritikai vizsgálata kiemelt jelentőségű. Pozitív szemléletben, ha a virtuális adatfelvétel módszertana kidolgozható, azáltal drasztikusan csökkenthető a valós adatgyűjtéshez szükséges idő és az erőforrások mértéke, ami a társadalomkutatás demokratizálódását és decentralizálódását jelentheti. Ezen gondolat szerint az eljárás a surveyadatok csökkenő validitásának problémájára is megoldást kínálhat. Azonban be kell azt is látni, hogy ez a probléma ennél összetettebb, a kritikai szemlélet elengedhetetlen ahhoz, hogy az LLM-ek ismert torzításainak hatásai ne maradjanak feltáratlanul, és még inkább ahhoz, hogy tudatosan figyelembe vegyünk az internet nyelvi terében nem reprezentált társadalmi valóságot.

A tanulmány a GPT (Generative Pre-training Transformer, az OpenAI által fejlesztett nagy nyelvmodell) által generált virtuális válaszadók és a European Social Survey adatfelvétel valós válaszadóinak válaszait hasonlítja össze statisztikai módszerekkel, a különböző promptolási technikák (a nyelvmodellnek megadott bemenetek meghatározása), valamint a modelleknek beadott változatos típusú és gazdagságú kontextuális információk hatását vizsgálva. Az eredmények azt mutatják, hogy a GPT élethű válaszokat generálhat, és előhívhatja a társadalmi mintákat a neki megadott korlátozott kontextuális információból, amennyiben megfelelő példákat adunk számára, azonban ezek nélkül nem teljesít elfogadhatóan. Mindezek alapján kétségek merülhetnek fel bennünk a GPT – mint virtuális válaszadó – sikeres szereplésével kapcsolatosan és minden olyan kutatással szemben, amely a GPT általános társadalomtudományi használatának sikerességéről számol be.

Kulcsszavak: nagy nyelvmodellek, virtuális válaszadók, GPT

\* A tanulmány a Kulturális és Innovációs Minisztérium ÚNKP-23-3 kódszámú Új Nemzeti Kiválóság Programjának a Nemzeti Kutatási, Fejlesztési és Innovációs Alapból finanszírozott szakmai támogatásával, illetve a Kulturális és Innovációs Minisztérium EKÖP-24 kódszámú Egyetemi Kiválósági Ösztöndíj Programjának a Nemzeti Kutatási, Fejlesztési és Innovációs Alapból finanszírozott szakmai támogatásával készült.

The emergence of large language models (LLMs) creates new opportunities for social research methods. For any research where language mediates empiricism, it becomes a realistic possibility to include virtual respondents simulated by LLMs as respondents instead of or in addition to real respondents.

The potential of artificial intelligence-generated data is such that methodological and critical analysis is of paramount importance. From a positivist perspective, if a methodology for virtual data collection can be developed, the time and resources required for real data collection can be drastically reduced, which could mean the democratisation and decentralisation of social research. The procedure could offer a solution to the problem of the decreasing validity of survey data. However, it must be acknowledged that this problem is more complex than that, a critical approach is essential to ensure that the effects of the known biases of LLMs do not remain unexplored, and even more so to consciously take into account the social reality not represented in the linguistic space of the Internet.

The study compares the responses of virtual respondents generated by GPT (Generative Pre-training Transformer, a large language model developed by OpenAI) and real respondents to the European Social Survey using statistical methods, examining the effects of different prompting techniques (defining inputs for the large language model) and the varying types and richness of contextual information input to the models. The results show that GPT can generate lifelike responses and elicit expected patterns from the limited contextual information it is given if it is provided with appropriate examples, but that it does not perform acceptably without them. All of this leads us to doubt the success of GPT as a virtual respondent and any research that reports on the success of GPT use in general social science applications.

Keywords: large language models, virtual respondents, GPT

A nagy mennyiségben rendelkezésre álló adatok természete, az adatgenerálás folyamatának jellemzői és az adatok forrása komoly kérdéseket vetnek fel az elemzők számára már a kutatás megkezdése előtt. A tanulmány a fenti kurrens és dinamikus bővülő szakterület egy kisebb egységével, a nagy nyelvmodellek által generált adatok társadalomtudományi használatának kritikai elemzésével foglalkozik.

A megalapozó modellek (*foundation models*) olyan generatív mesterséges intelligencia (MI) -eszközök, amelyek tanulóadatai lényegileg a teljes internetet lefedik (*Touvron et al., 2023*), ezért azokat specifikus problémák megoldásához sem kell lényegileg módosítani, pl. új adatokon (tovább) tanítani. Az ilyen mélytanuláson (*deep learning*) alapuló nagy nyelvmodellekre (*Large Language Model, LLM*) épülő mesterséges intelligencia (*artificial intelligence, AI*) megjelenése egy új – eddig kiaknázatlan – lehetőséget teremtett a társadalomkutatási megismerés módszerei szempontjából. Mind a kvalitatív, mind a kvantitatív empirikus kutatások esetében, amelyeknél a nyelv közvetíti az empiriát, reális lehetőséggé vált valós személyek – mint kvalitatív vagy kvantitatív adatközlők – helyett vagy mellett az LLM-ek – mint adatszolgáltatók – által generált virtuális személyek bevonása,

virtuális válaszok generálása. Az így létrehozott adatokat virtuális mintának vagy másképpen szilíciummintának (*Argyle et al., 2022*) nevezhetjük.

*Argyle és szerzőtársai (2022, 2023)* kimutatták, hogy az American National Election Studies közvéleménykutatás (*ANES, 2021*) esetében a GPT-3-modell (*Brown et al., 2020*) által generált válaszok átmennek az ún. társadalomtudományos Turing-teszten, azaz a kutatók nem tudtak különbséget tenni a valós személyek és a szimulált, fiktív személyek válaszai között, továbbá ezen válaszok konzisztensek a virtuális személy kutatók által megadott szociodemográfiai jellemzőivel. Általánosan is elmondható, hogy az ilyen irányba induló vizsgálatok, például néhány szociálpszichológiai kísérlet (pl. Milgram-kísérlet) virtuális ágenssel való megismétlése és politikai véleménykutatásokban való tesztelése (*Aher et al., 2023*) arra enged következtetni, hogy a nagy nyelvmodellek számtalanszor demonstrált nyelvi gazdagsága képes a valós emberi válaszok és reakciók hű leképzésére.

Ugyanakkor a modellek által generált adatok minősége nagymértékben függ azok kinyerésének módjától. Emiatt mára a prompt mérnökségnek (*prompt engineering*), azaz a kívánt kimenetekhez (outputokhoz) szükséges legjobb bemenetek (inputok) megkeresésének önálló – bár egyelőre még nem kanonizált – módszertana alakult ki (*Yao et al., 2023*). A társadalomkutatási céllal használható adatok kinyeréséhez kritikus kérdés, hogy milyen prompt (modellnek megadott parancs) segítségével definiálható a releváns válaszokhoz szükséges kontextus.

Általánosságban is elmondható, hogy a mesterséges intelligencia által generált adatokban rejlő potenciál olyan mértékű, hogy annak módszertani és kritikai vizsgálata kiemelt jelentőségű. Pozitívista szemléletben, ha a virtuális adatfelvétel módszertana kidolgozható, azzal a valós adatfelvételre fordított idő és erőforrásigény drasztikusan csökkenthető, ami a társadalomkutatás demokratizálódását, decentralizálását jelentheti, de legalább is lehetőséget arra, hogy az eddigiekhez képest sokkal általánosabb virtuális pilotkutatásokkal lehessen felkészülni a tényleges adatfelvételekre, és szélesebb tere legyen ezen adatfelvételek javításának (pl. adatok pótlásával), kiegészítésének. Ezen gondolat szerint az eljárás a surveyadatok – alacsony válaszadási arány miatti – csökkenő validitásának problémájára is megoldást kínálhat. Elengedhetetlen ugyanakkor a kritikai szemlélet ahhoz, hogy a nagy nyelvmodellek már ismert torzításainak (*Schramowski et al., 2022*) hatásai ezen alkalmazási lehetőségeknél se maradjanak feltáratlanok, illetve még inkább az internet nyelvi terében le nem képződő társadalmi valóság tudatos figyelembevételéhez. Bizonyos csoportok például kevésbé férnek hozzá az digitális eszközökhöz és az internethez, kevésbé érik el az online teret, vagy kevésbé tudják használni azt, tehát egyes csoportok láthatósága sokkal alacsonyabb, ami azt is jelenti, hogy az őket érintő tartalmak kevésbé lesznek reprezentálva a nyelvi modellekben.

Jól dokumentált továbbá az LLM-ekre épülő interaktív mesterséges intelligencia-applikációk – pl. a ChatGPT csevegőfelületének – felhasználók általi antropomorfizálása, amelyek túlzott bizalmat keltenek a modellek kimenetében azáltal, hogy intelligens ágensként tüntetik azokat fel (*Heersmink, 2024*). *Shanahan (2023)* szerint a látszólagos ágenciából és az ún. hallucinációk – tehát a modellek által generált válaszok tényszerűnek tűnő, de valójában fiktív tartalmi – generálta problémák hatása csökkenthető, amennyiben az LLM-ekről és általában a mesterséges intelligenciáról folytatott diskurzusban tudatosan kerüljük a „tudja”, „hiszi”, „gondolja” stb. antropomorfizáló kifejezéseket. *Mahowald és szerzőtársai (2024)* szerint az „ágensjellegű szemlélet” helyett az „eszközjellegű szemlélet” alkalmazása, továbbá az LLM-ek „formális” – tehát a nyelvi szerkezetekre és szabályszerűségekre vonatkozó – nyelvi kompetenciájának és „funkcionális” kompetenciájának megkülönböztetése segít abban, hogy a felhasználók realisabban mérjék fel az ilyen rendszerek lehetőségeit és korlátait. A bemutatott kiértékelési eljárás során megpróbáltam érvényesíteni a fenti figyelmeztetések tanulságait.

## 1. Adatok és módszerek

A kutatás nagy nyelvmodellek (GPT-variánsok) által generált virtuális válaszadók és a European Social Survey (*ESS, 2022*) nemzetközi összehasonlító adatfelvétel valós válaszadóinak válaszait hasonlítja össze statisztikai módszerekkel, vizsgálva a különböző promptolási technikák, valamint a modelleknek beadott változatos típusú és gazdagságú kontextuális információk hatását.

Az elemzés első lépésében az ESS 10. hullámának Magyarországra és az Egyesült Királyságra vonatkozó almintáira, a GPT két variánsa – GPT-3.5-turbo, GPT-4 – által generáltam virtuális válaszadókat és válaszokat a demokráciával való elégedettséget mérő kérdőívkérdésre. Második lépésben az így kapott szilíciumminta és a valós minta eredményeinek statisztikai összevetését végeztem el mind a két (magyar és brit) almintára vonatkozóan. Az összehasonlításnál a GPT két modellvariánsának teljesítménybeli különbségeit, a különböző promptolási technikákat, a megadott bemenetek közötti különbségeket is vizsgáltam. Különböző modellbeállítások és bemenetek mellett a modelleknek megadott változatos típusú és gazdagságú kontextuális információk hatását is elemeztem, például a *zero-shot*, illetve a *few-shot* eseteket. Az előbbi azokra az esetekre utal, amikor bemenetben nem szerepelt példa a modell számára a döntés meghozatalához, az

utóbbi pedig azokra utal, ahol több példa is szerepelt. A példák típusait is megkülönböztettem, így a véletlen példákat, illetve egy adott virtuális válaszadó számára leginkább releváns, célzott példákat és azok szerepét is elemezni tudtam a modellteljesítmények értékelésekor. A virtuális válaszadó számára releváns példákhoz az ESS-almintákban szereplő valós – a virtuális válaszadóhoz hasonló – valódi személyek közül választottam az egymásra leginkább hasonlítókat. Harmadik lépésben az összehasonlítás eredményeinek kiértékelése és a teljesítmények validálása zajlott. A kutatás ezen szakasza magában foglalta a modellek – specifikus értékelési metrikák szerinti – algoritmikus hűségének mérését is: a választott kérdőív kérdéseinek mérési szintje diktálta az összehasonlítási metrikák kiválasztását, például a válaszok eloszlását, a korrelációs mintázatokat, és azt, hogy a változók közötti összefüggések megismétlődtek-e (ezekről még részletesebben a későbbi bekezdésekben lesz szó).

A tanulmányban kizárólag egyetlen kérdés elemzését mutatom be részletesen, azt, amely a demokráciával való elégedettséget méri egy 11 fokú skálán. A kapcsolódó változó neve „stfdem”, értékei pedig 0 és 10 között lehetnek, ahol a 0 azt jelenti, hogy „Teljes mértékben elégedetlen”, és a 10 pedig azt, hogy „Teljes mértékben elégedett” a válaszadó. A kérdőív kérdés magyarul így hangzott: „Mennyire elégedett a demokrácia működésével az országban?”, angolul pedig: „*How satisfied with the way democracy works in country?*”; 0: „*Extremely dissatisfied*”, 10: „*Extremely satisfied*”.

Az elemzés előkészítésként, a promptok felépítéséhez, valamint a generált válaszok korrelációinak vizsgálatához kiválasztottam 5 változót az ESS-ben rendelkezésre állók közül, amelyek szociológiai szempontból relevánsan magyarázhatják a célváltozót. A következő változókat választottam ki a politika iránti érdeklődés magyarázatához: nem, életkor, legmagasabb iskolai végzettség, bal-jobb skála szerinti politikai attitűd és a jogrendszerbe vetett bizalom mértéke. Ahogy az 1. táblázatban látható, nem minden esetben van szignifikáns kapcsolat az adott modellben egy kiválasztott magyarázóváltozó és a célváltozó között, így az is vizsgálható, hogy az LLM-ek válaszaik korrelálnak-e olyan változókkal, amelyek ténylegesen nem függenek össze a célváltozóval.

1. táblázat

**A Magyarországra és az Egyesült Királyságra vonatkozó  
ESS-mintákra számolt regressziós modellek mutatói**  
*Regression model outputs calculated for the ESS samples of  
Hungary and the United Kingdom*

	Becsült együttható	Standard hiba	<i>t</i> -érték	<i>p</i> -érték	Standardi- zált, becsült együttható
ESS magyar minta					
Konstans	0,733	0,115	6,367	< 0,001	nem releváns
Nem (referencia: nő)	0,048	0,040	1,182	0,237	0,020
Életkor	0,002	0,001	1,444	0,149	0,025
Iskolai végzettség	-0,025	0,026	-0,966	0,334	-0,017
Bal-jobb skálán mért beállítódás	0,147	0,009	16,092	< 0,001	0,294
Jogrendszerbe vetett bizalom	0,245	0,009	28,034	< 0,001	0,514
R-négyzet értéke	0,4582				
F-statisztika értéke	311,8				
F-statisztika szabadságfokai	5 és 1843				
F-statisztika <i>p</i> -értéke	< 0,001				
ESS brit minta					
Konstans	1,238	0,175	7,095	< 0,001	nem releváns
Nem (referencia: nő)	-0,113	0,060	-1,886	0,060	-0,049
Életkor	0,002	0,002	1,314	0,189	0,036
Iskolai végzettség	-0,040	0,024	-1,694	0,090	-0,046
Bal-jobb skálán mért beállítódás	0,136	0,015	9,359	< 0,001	0,251
Jogrendszerbe vetett bizalom	0,195	0,013	15,449	< 0,001	0,407
R-négyzet értéke	0,2429				
F-statisztika értéke	73,34				
F-statisztika szabadságfokai	5 és 1143				
F-statisztika <i>p</i> -értéke	< 0,001				

Forrás: az ESS-adatfelvétel magyar és brit adatainak elemzése alapján saját elemzés és szerkesztés.

A modellek számára a fenti változókat felhasználva virtuális válaszadókat határoztam meg, amelyek szerepébe helyezkedve válaszoltak a kérdésre. A bemenetként megadott információk például a következőképpen kerültek meghatározásra angolul (a magyar fordítás a lábjegyzetben<sup>1</sup>):

„This is the task: Pretend you are a Hungarian 64-year-old female with vocational education who is slightly right leaning politically, and who has a lot of distrust in the country’s legal system. Question: On a scale of zero to ten, how satisfied are you with the way democracy works in your country? Zero means extremely dissatisfied, ten means you are extremely satisfied with the way democracy works in your country. Please consider the complete scale, not just zero and ten. Please answer the question!”

A GPT-3.5-turbo válasza pedig a következő volt a fenti promptra (a magyar fordítás a lábjegyzetben<sup>2</sup>):

„As a 64-year-old Hungarian female with vocational education, and a slightly right-leaning political outlook with some distrust in the legal system, you might view democracy in your country through a nuanced lens. Given these factors, your level of satisfaction with the way democracy works in Hungary could be 6 on a scale of 0 to 10. While you may appreciate certain aspects of democratic processes, such as the right to vote and freedom of expression, your concerns about aspects of the legal system and possibly political dynamics may lead to a moderate level of satisfaction.”

A nagy nyelvmodelleknek megadott promptok és a példák típusának hatását is ellenőriztük azáltal, hogy teszteltem a *zero-shot* tanulást, amelyben kizárólag a virtuális személy karakterjegyeit (a fentiek szerint) határoztam meg, de nem adtam példát a válaszokra, továbbá a *few-shot* vagy általánosabban *n-shot* tanulást is. Az utóbbi úgy történik, hogy a modellnek *n* darab példát adunk egy adott feladat megoldására a promptban. Ennek a megközelítésnek az erősségét *Brown és szerzőtársai (2020)* eredeti GPT-3 tanulmánya is kiemelte. (Más teljesítményt nyújtó promptolási technikákról lásd *Wei és szerzőtársai {2023}*, illetve *Yao és szerzőtársai {2023}* munkáját.) Az *n-shot* tanulás esetében 5 példa beadását választottam.

<sup>1</sup> „Ez a feladat: Tegyük fel, hogy Ön egy 64 éves, szakképzett magyar nő, aki politikailag enyhén jobboldali beállítottságú, és aki nagyon bizalmatlan az ország jogrendszerével szemben. Kérdés: Egy 0-tól 10-ig terjedő skálán mennyire elégedett azzal, ahogyan a demokrácia működik az Ön országában? A 0 azt jelenti, hogy rendkívül elégedetlen, a 10 azt, hogy rendkívül elégedett azzal, ahogyan a demokrácia működik az Ön országában. Kérjük, ne csak a 0-t és a 10-et, hanem a teljes skálát vegye figyelembe! Kérem, válaszoljon a kérdésre!”

<sup>2</sup> „64 éves magyar nőként, szakképzettséggel, enyhén jobboldali politikai beállítottsággal és némi bizalmatlansággal a jogrendszerrel szemben, talán árnyaltan szemléli a demokráciát az országában. Ezeket a tényezőket figyelembe véve az Ön elégedettségi szintje a demokrácia magyarországi működésével kapcsolatban egy 0-tól 10-ig terjedő skálán 6-os lehet. Miközben Ön értékelheti a demokratikus folyamatok bizonyos aspektusait, például a választójogot és a véleménynyilvánítás szabadságát, a jogrendszer egyes aspektusaival és esetleg a politikai dinamikával kapcsolatos aggályai mérsékelt elégedettségi szinthez vezethetnek.”



Ezen *5-shot* tanuláshoz az alábbi két esetét is vizsgáltam: 1. véletlen példák biztosítása, ahol a véletlenszerűen választott valós személyek és válaszaik képezték a bemenetként megadott adatokat; 2. célzott példák megadása, ahol az adathalmazban meghatározott jellemzőkkel való hasonlóság alapján választottam ki a virtuális válaszadóhoz legközelebbi valós válaszadók releváns válaszait példaként. Ez nem része az itt tárgyalt kiértékelésnek, de egy bővebb vizsgálatban érdemes lenne mérni a promptban megadott virtuális válaszadóhoz felhasznált változók körének szerepét. Vélhetően több változó használata, azaz egy gazdagabb kontextus megadása javítja a válaszok minőségét, továbbá az is számíthat, hogy a változók együttesen ténylegesen milyen mértékben határozzák meg a célváltozót.

A modellek tanításakor a nyelvek kiegyensúlyozatlan reprezentációjának lehetséges hatását mérsékelve úgy döntöttem, hogy a promptoláshoz az angol nyelvet használom egységesen. Ezáltal a modellek nyelvtudásbeli különbségeit kontrollálni lehet, amíg az ország explicit megadása lehetővé teszi a modellek számára, hogy a bennük lévő országspecifikus információkat figyelembe vegyék.

Az LLM-ek összehasonlításához a különböző promptolási technikákkal és beállításokkal kétféle értékeltési módszert alkalmaztam. Először kiszámítottam a Kullback–Leibler (KL)-eltérést a célváltozó valós értékeiből adódó és a szilícium-mintákból adódó eloszlásai között. A KL-becslések eloszlásait bootstrapmódszerrel (minden esetben 2000 elemű bootstrapmintákkal) közelítettem, hogy megbecsüljem azok mintavételi ingadozását. Az első kiértékelési kör tehát az eloszlás-szintű hűségre helyezi a hangsúlyt, figyelmen kívül hagyva, hogy a generált egyéni szintű válaszok megfelelnek-e a valós válaszoknak. A KL-divergenciát azért választottam, mert ez a valószínűségi eloszlások közötti különbség mindenütt elterjedt, komoly információelméleti háttérrel rendelkező mértékegysége (*Garrido, 2009*), és a modelltől származó eloszlásnak a valódi eloszlás helyett történő felhasználásával elvesztett információ mértékeként értelmezhető.

Másodszor, a demokráciával való elégedettségre vonatkozó regressziós modellt a nem (férfi vagy nő), az életkor (folytonos változó), az iskolai végzettség (4 kategóriás változó, értékei maximum általános iskolai végzettség, szakma, középfokú és felsőfokú végzettség), a bal-jobb skála szerinti politikai attitűd (Likert-skála, 0-tól 10-ig terjedő skálán mérve, balról jobbra haladva), illetve a jogrendszerbe vetett bizalom mértéke (Likert-skála, 0-tól 10-ig terjedő skálán mérve, az „Egyáltalán nem bízok” kategóriától a „Teljes mértékben megbízok a jogrendszerben” kategóriáig) magyarázóváltozókkal illesztettem, majd összehasonlítottam a generált értékekből kapott standardizált regressziós együtthatókat a valós minta alapján becsültekkel. A második kiértékelési kör tehát az egyéni szintű hűségre támaszkodik, mivel a regressziós modell az egyén által összekötött változók értékei alapján ragadja meg a korrelációs struktúrát. Ezt követően egy gyakran alkalmazott dimenziócsökkentő eljárás (UMAP) használatával az öt magyarázóváltozó

standardizált regressziós együtthatóinak vizualizációját is elvégeztem a különböző beállításokra vonatkozóan, hogy az eredmények vizuálisan is jól összevethetőek legyenek. A következő fejezetben a fent említett két összehasonlítás eredményeit mutatom be.

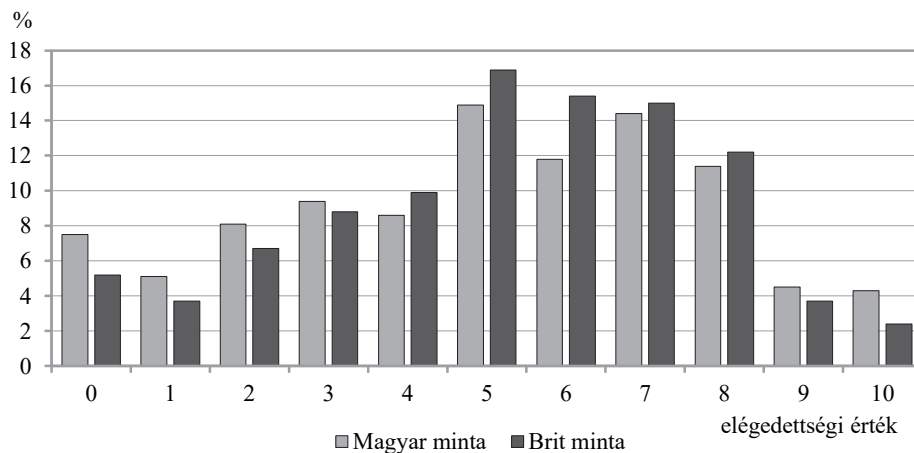
## 2. Eredmények

### 2.1. A demokráciával való elégedettség

Az eredmények ismertetése előtt röviden bemutatom a demokráciával való elégedettség változó eloszlását a valós ESS-mintákban, országonként, a Magyarországra és az Egyesült Királyságra vonatkozó almintákban.

1. ábra

**A demokráciával való elégedettség változó kategóriáinak eloszlása**  
*Distribution of the variable satisfaction with democracy*



Forrás: az ESS-adatfelvétel magyar és brit adatainak elemzése alapján készült saját elemzés és szerkesztés.

Az 1. ábrán a demokráciával való elégedettség válaszkategóriák szerinti százalékos eloszlása látható az ESS 10. hullámának adatain országonként. A válaszkategóriák közül a 0 jelenti azt, hogy „Teljes mértékben elégedetlen”, a 10 pedig azt, hogy „Teljes mértékben elégedett”. A magyar és a brit társadalomban is az 5-ös érték (a középérték) a leggyakoribb. A magyar mintán a teljes mértékben elégedetlenek aránya 7,5, a teljes mértékben elégedetteké pedig 4,3%. A köztes negatív

attitűdű (az 1–4-es) értékek arányai rendre 5,1; 8,1; 9,4; 8,6%, illetve pozitív attitűdű (a 6–9 közötti) értékek arányai pedig 11,8; 14,4; 11,4; 4,5%. A brit minta esetében valamelyest különbözik az eloszlás: a teljes mértékben elégedetlenek (0-s érték) és az inkább elégedetlenek (1–3) alacsonyabb arányban szerepelnek, míg a kissé elégedetlen, semleges és inkább elégedett (4–8-as értékkel) válaszolók pedig magasabb arányban szerepelnek, érdekes továbbá, hogy a nagyon elégedettek (9) és a teljes mértékben elégedettek (10) aránya szintén alacsonyabb, mint a magyar mintán látható, rendre 3,7 és 2,4%.

## 2.2. Kullback–Leibler-divergenciák elemzése

A valós és a szilíciumminták közötti Kullback–Leibler (KL)-divergencia vizsgálata jelentős különbségeket mutatott az GPT-variánsok és a különböző beállítások tekintetében, lásd az 2. ábrát Magyarországra, a 3. ábrát az Egyesült Királyságra vonatkozóan. A KL-bootstrap-eloszlásokat a szemléltetés érdekében simítottuk, de az értékek mindig nem negatívak. Az ábrákon az egyszerűség kedvéért a modellvariánsok és beállítások rövid neve szerepel.

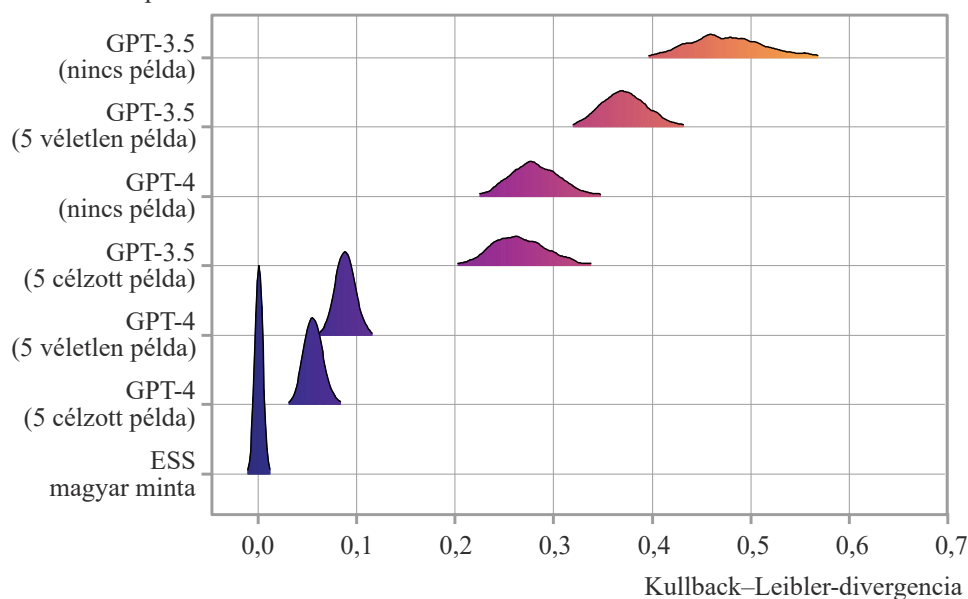
A magyar almintára (2. ábra) számolt a Kullback–Leibler-divergencia szerint a GPT-4 5 célzott példával teljesített a legjobban, az generálta az eredetihez legközelebbi eloszlást, ezt követően a GPT-4 5 véletlen példával következett, de nincs szignifikáns eltérés ( $p = 0,015$ ) KL-divergencia szerint. Az eltérések szignifikanciája szerint lényegileg három csoport különböztethető meg: a GPT-4 célzott és véletlen példákkal teljesít a legjobban, ezeknél szignifikánsan gyengébbek, de egymáshoz hasonló teljesítményűek a GPT-3.5-turbo és a GPT-4 példák nélkül, végül leggyengébbek a GPT-3.5-turbo véletlen példákkal és példák nélkül promptolva. Összességében a magyar adatokra az 5 célzott példával promptolt modellek teljesítettek jobban minden modelltípus esetén, illetve a GPT-4 jobb eredményeket ért el, mint a GPT-3.5-turbo modell.

A 2. ábra a magyar ESS-alminta és a GPT-variánsok által generált szilíciumminták között számolt Kullback–Leibler-divergencia bootstrapmintáinak sűrűségfüggvényeit ábrázolja a demokráciával való elégedettségre vonatkozóan a különböző modelltípusok és -beállítások szerint. A vízszintes tengely: Kullback–Leibler-divergencia, a függőleges tengely pedig a becült sűrűségfüggvény értéke modelltípusonként és promptolási technikák szerint.

2. ábra

**Modelltípusok és -beállítások szerinti Kullback–Leibler-divergencia  
a magyar ESS-alminta és a GPT-variánsok által generált szilíciumminták között**  
*Kullback-Leibler divergence by model type and -setting between  
the Hungarian ESS sample and the silicon samples generated by GPT-variants*

Modelltípus és -beállítás



Forrás: az ESS-adatfelvétel magyar adatainak, illetve a kutatáshoz használt nagy nyelvmodellek által generált adatok statisztikai elemzése alapján saját elemzés és szerkesztés.

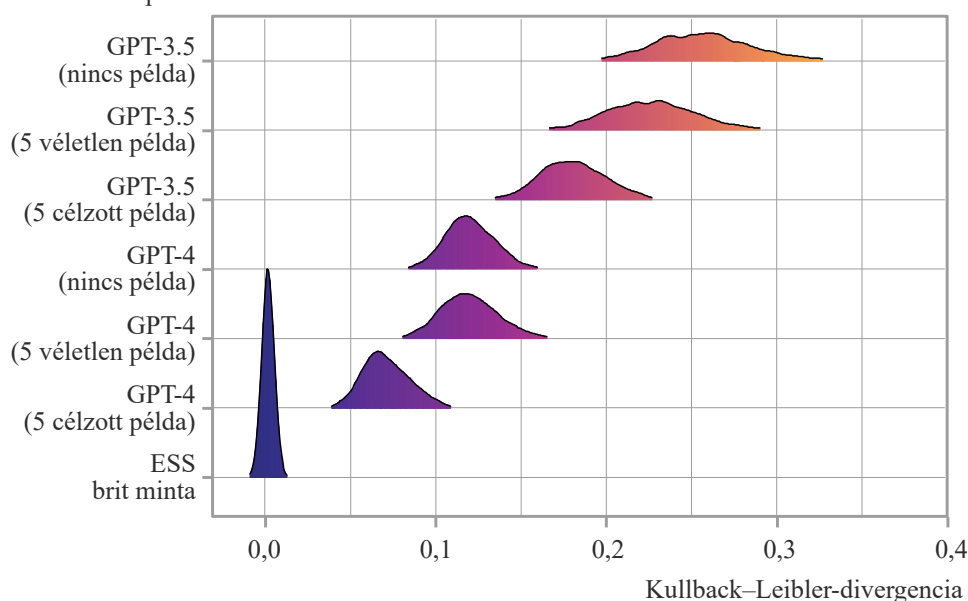
A magyar almintára (2. ábra) számolt a Kullback–Leibler-divergencia szerint a GPT-4 5 célzott példával teljesített a legjobban, az generálta az eredetihez legközelebbi eloszlást, ezt követte a GPT-4 5 véletlen példával. Az eredmények alapján ezeket követően szerepel a GPT-3.5 modellje 5 célzott példával, majd a GPT-4 példák nélkül, és végül a GPT-3.5 modelljei 5 véletlen példával és példák nélkül.

A 3. ábra a brit ESS-alminta és a GPT-variánsok által generált szilíciumminták között számolt Kullback–Leibler-divergencia bootstrapmintáinak sűrűségfüggvényeit ábrázolja a demokráciával való elégedettségére vonatkozóan a különböző modelltípusok és -beállítások szerint. A vízszintes tengely Kullback–Leibler-divergencia, a függőleges tengely pedig a becsült sűrűségfüggvény értéke modelltípusonként és promptolási technikák szerint.

3. ábra

**Modelltípusok és -beállítások szerinti Kullback–Leibler-divergencia  
a brit ESS-almintára és a GPT-variánsok által generált szilíciumminták között**  
*Kullback-Leibler divergence by model type and -setting between  
the British ESS sample and the silicon samples generated by GPT-variants*

Modelltípus és -beállítás



Forrás: az ESS-adatfelvétel brit adatainak, illetve a kutatáshoz használt nagy nyelvmodellek által generált adatok statisztikai elemzése alapján saját elemzés és szerkesztés.

A brit almintára (3. ábra) számolt a Kullback–Leibler-divergencia szerint – hasonlóan a magyar eredményekhez – a GPT-4 5 célzott példával teljesített a legjobban, az generálta az eredetihez legközelebbi eloszlást, ezt követte a GPT-4 5 véletlen példával. A brit ESS-adatokra vonatkozóan is elmondható, hogy általában az 5 célzott példával promptolt modellek teljesítettek jobban minden modell típus esetén, illetve, hogy összességében a GPT-4 hozott jobb eredményeket. Az is leolvasható az ábráról, hogy kisebb KL-divergenciák adódtak, azaz a modellek által generált adatok eloszlása általánosan jobban hasonlított a valóshoz, mint ami a magyar adatoknál megfigyelhető.

A magyar és a brit almintára vonatkozó Kullback–Leibler-divergenciák modellek és beállítások szerinti eltéréseinek szignifikanciáit tartalmazó táblázatokat lásd a Függelékben (F1. és F2. táblázat).

### 2.3. Regressziós modellek elemzése

A különböző modelltípusok és -beállítások regressziós együtthatói is különbségeket mutattak a szilíciumminták teljesítményében. A 2. táblázat a két alminta valós adataira vonatkozó regressziós eredményeket mutatja. Érdeemes megnézni a bal-jobb skálán mért politikai attitűd és a jogrendszerbe vetett bizalom mértékei változókat, mert ezek esetén kontroll alatt tartva a szociodemográfiai változókat, szignifikáns hatásokat látunk. Általánosan elmondható mind a két ország esetében, hogy minél jobboldalibb a válaszadó, illetve minél jobban bízik a jogrendszerben, annál elégedettebb a demokráciával.

2. táblázat

**Regressziós együtthatók Magyarország és az Egyesült Királyság ESS-almintáira, a valós és a legjobban teljesítő modell által generált adatokra vonatkozóan**  
*Regression coefficients for the ESS samples of Hungary and the United Kingdom, for the real and best performing model generated data*

Magyarázóváltozók/minták	ESS			
	magyar minta		brit minta	
	valós	GPT-4 5 példa	valós	GPT-4 5 példa
Nem	0,048	0,038	-0,081	-0,056
Életkor	0,002	0,000	0,002	0,003
Iskolai végzettség	-0,025	-0,019	-0,036	0,003
Bal-jobb skálán mért politikai attitűd	0,147	0,138	0,118	0,101
Jogrendszerbe vetett bizalom mértéke	0,245	0,242	0,178	0,190

Forrás: az ESS-adatfelvétel magyar és brit adatainak, illetve a kutatáshoz használt nagy nyelvmodellek által generált adatok statisztikai elemzése alapján saját elemzés és szerkesztés.

A modellek a demokráciával való elégedettséget – amely egy 11 fokú skálán mért, 0 és 10 közötti értéket felvevő változó – kvázi folytonos kimenetként jelzik előre, a magyarázóváltozók pedig a nem, az életkor, az iskolai végzettség (0-tól 4-ig, szintén kvázi folytonos változó) és az önbesorolás a bal-jobb skálán (Likert-skála, 0-tól 10-ig terjedő skálán mérve), illetve a jogrendszerbe vetett bizalom (Likert-skála, 0-tól 10-ig terjedő skálán mérve).

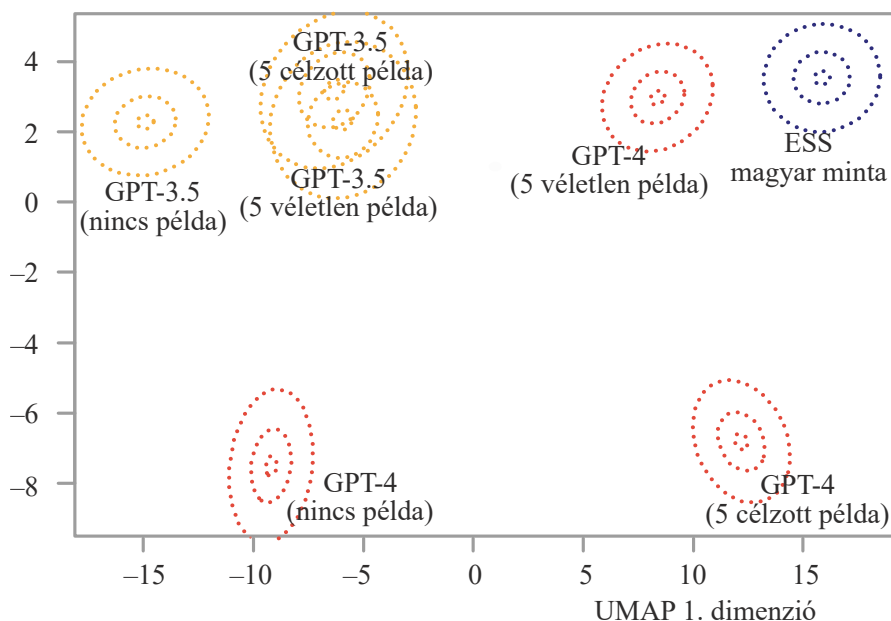
A 4–5. ábra a kiválasztott európai országok regressziós együttható vektorainak bootstrapeloszlásaira vonatkozó UMAP-projekcióit mutatják, amelyekből kiderül, hogy nincs olyan általánosan legjobb modell, amely mindkét alminta esetében egyformán jól reprodukálná az összes együtthatót.

4. ábra

**A magyar ESS-minta és a GPT-variánsok által különböző promptolási technikák mellett generált szilíciummintákból becsült együtthatók összehasonlítása**

*Comparison of coefficients estimated from Hungarian ESS sample and silicon samples generated by GPT-variants under different prompting techniques*

UMAP 2. dimenzió



Forrás: az ESS-adatfelvétel magyar adatainak, illetve a kutatáshoz használt nagy nyelvmodellek által generált adatok statisztikai elemzése alapján saját elemzés és szerkesztés.

A 4. ábrán a Magyarország ESS-almintájából becsült regressziós együtthatók összehasonlítására kerül sor, a vizsgált GPT-variánsok által különböző promptolási technikák mellett generált szilíciummintákból becsült együtthatókkal. Az ábra az együtthatóvektorok eloszlásainak UMAP-dimenziócsökkentéssel kapott két dimenzióra vetített képét mutatja. A szaggatott ellipszisek az együtthatóvektorok bootstrapeloszlásának percentiliseit (0,05, 0,55 és 0,95) jelölik.

A magyar alminta esetében (4. ábra) a GPT-4 5 véletlen példával teljesített a legjobban. Ezen elemzés koncepcióját tekintve különbözött a korábitól, emiatt nem feltétlenül meglepő, hogy teljesítményük alapján más sorrendje adódik a modell típusoknak és -beállításoknak. Az azonban ennél az elemzésnél is elmondható, hogy az 5 (akár véletlen, akár célzott) példával ellátott modellek közelebbi eredményeket hoztak a valósakhoz, a példa nélküli modellek (*0-shot*) rosszabbul teljesítettek.

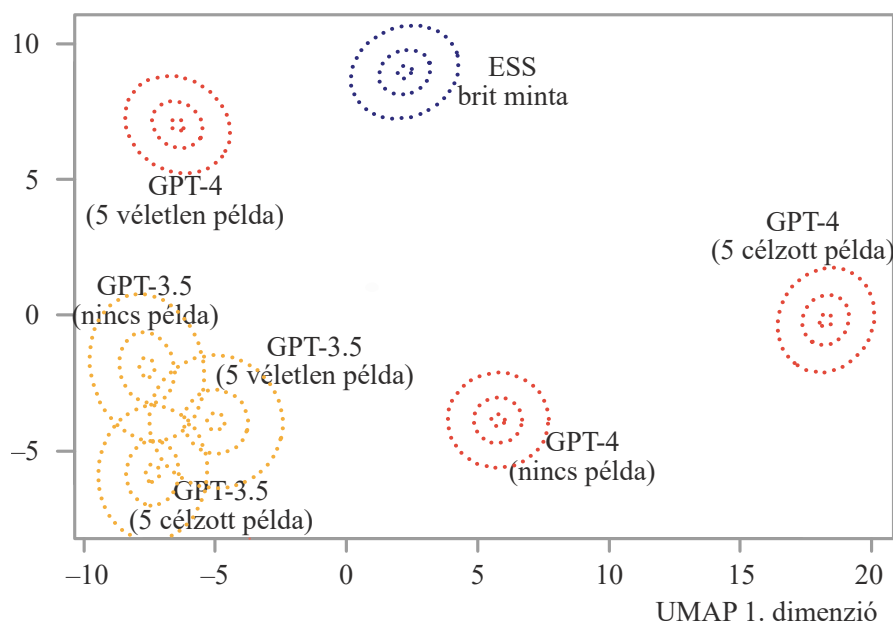
A brit alminta esetében (5. ábra) is a GPT-4 5 véletlen példával teljesített a legjobban. Az egyes modell típusok és -beállítások egymáshoz való viszonya is hasonló a két almintán. A GPT-3.5-turbo modellel kapott regressziós együtthatók lényegileg azonosak függetlenül a példák számától és minőségétől. A GPT-4 esetében nagyobb eltérések adódtak, azaz érdemi szerepe volt a prompt konstrukciójának.

5. ábra

**A brit ESS-minta és a GPT-variánsok által különböző promptolási technikák mellett generált szilíciummintákból becsült együtthatók összehasonlítása**

*Comparison of coefficients estimated from British ESS sample and silicon samples generated by GPT-variants under different prompting techniques*

UMAP 2. dimenzió



Forrás: az ESS-adatfelvétel brit adatainak, illetve a kutatáshoz használt nagy nyelvmodellek által generált adatok statisztikai elemzése alapján saját elemzés és szerkesztés.

Az 5. ábrán az Egyesült Királyság ESS-almintájából becsült regressziós együtthatók összehasonlítására kerül sor a vizsgált GPT-variánsok által különböző promptolási technikák mellett generált szilíciummintákból becsült együtthatókkal. Az ábra az együtthatóvektorok eloszlásainak UMAP-dimenziócsökkentéssel kapott két dimenzióra vetített képét mutatja. A szaggatott ellipszisek az együtthatóvektorok bootstrapeloszlásának percentiliseit (0,05; 0,55 és 0,95) jelölik.



## 2.4. A modellek hibáinak mintázatai

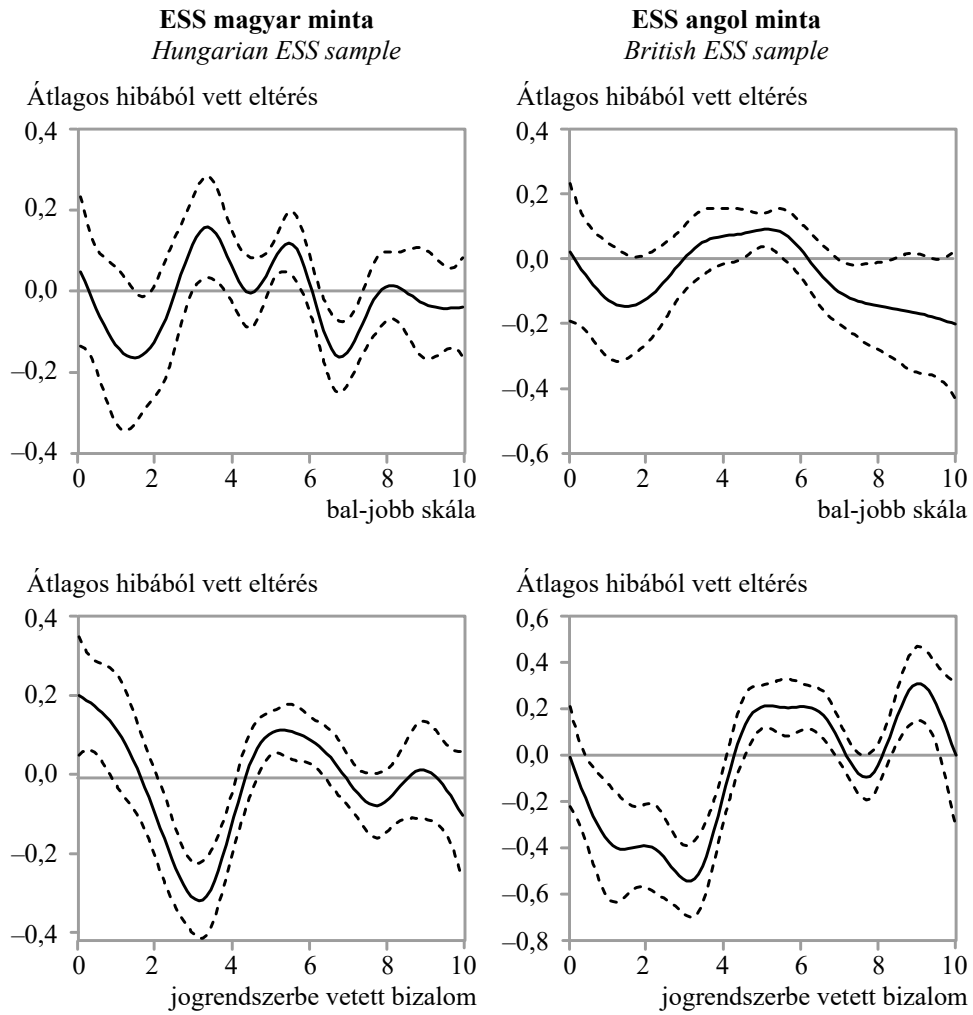
Annak érdekében, hogy az LLM-ek által generált és a valós értékek közötti eltérések tendenciáit vizsgáljam, a hibára – tehát a célváltozóra vonatkozóan a generált és a valós érték közötti különbségre – általánosított regressziós modelleket (GAM) illesztettem, amelyekkel nemlineáris összefüggések is modellezhetők azáltal, hogy a magyarázóváltozók *spline*-jaival becsli a modell kimeneti változójának feltételes várható értékét. A 6. ábrán látható a magyar és a brit mintára az 5 közeli példával promptolt GPT-4 modell hibáinak összefüggése a bal-jobb skálával és a jogrendszerbe vetett bizalommal. Ezek mellett mindkét regressziós modellben szerepelt a nem, a kor és az iskolai végzettség. Utóbbi háromnak nem volt szignifikáns összefüggése a modell hibájával.

A 6. ábrán vizsgált hiba a demokráciával való elégedettség célváltozón a modell általi és a valós válaszok különbsége, így a negatív hibák azt mutatják, hogy a modell alulbecsüli a tényleges értéket, a pozitív hibák pedig felülbecslést mutatnak. Az összefüggést leíró függvények a magyar és a brit hibákra illesztett általánosított additív modellekből származó regularizált köbös *spline*-ok. A regressziós modellek teljes outputja a Függelékben látható.

A magyar almintá esetében a GPT-4 modell hibája a bal-jobb skála alacsony értékeinél negatív, tehát azon válaszadók esetében, akik inkább baloldaliak, a modell egy kissé alacsonyabb elégedettséget mutat a demokráciával, mint a tényleges érték. Hasonló összefüggés látszik a jogrendszerbe vetett bizalom esetében: akiknek alacsony a jogrendszerbe vetett bizalma, azoknál a modell alacsonyabb elégedettséget becsült a demokráciával, mint a tényleges érték. A brit mintán megismétlődnek az előző tendenciák, de itt a két magyarázóváltozó magasabb értékei esetében is látható torzítás. Azon brit válaszadók esetében, akik a bal-jobb skála jobb szélére esnek, alacsonyabb a GPT-4 szerinti demokráciával való elégedettséget, mint a valós, valamint azoknál, akinek magas a jogrendszerbe vetett bizalma, a modell jellemzően magasabb elégedettséget mutat a demokráciával, mint a tényleges érték. A brit mintára vonatkozó eredmények összességében erősebb és egyértelműbb torzítási tendenciákat mutatnak, mint a magyarra vonatkozók. Az, hogy a bal-jobb skála szerint szélsőségesebb politikai beállítottságú személyek esetében a demokráciával a ténylegesnél alacsonyabb elégedettséget mutat a modell, egyfajta sztereotipikus mintázatnak tekinthető, ahogyan az is, hogy a jogrendszerben kevésbé bízók esetében a valósánál alacsonyabb, a jobban bízóknál pedig magasabb a demokráciával való elégedettsége a modell szerint.

6. ábra

**A GPT-4 modell (5 közeli, célzott példával) hibájának összefüggései a promptban megadott bal-jobb skála és a jogrendszerbe vetett bizalom változókkal**  
*The relationship of the error of the GPT-4 model (5 close, targeted examples) with the left-right scale and the trust in the legal system variables*



Forrás: a GPT-4 modell által – az ESS magyar és brit mintái alapján – generált adatokon végzett statisztikai elemzés alapján saját elemzés és szerkesztés.

### 3. Összefoglalás

A tanulmány a nagy nyelvmodellek társadalomtudományi kutatásokban való lehetséges felhasználásának feltárását tűzte ki célul, hogy hozzájárulhasson azok tudományos alkalmazási lehetőségeinek és korlátainak megismeréséhez. A GPT-variánsokat (GPT-3.5-turbo és GPT-4) és különböző promptolási technikákat vizsgáltam, összevetve az ESS-adatbázis 10. hullámának Magyarországra és az Egyesült Királyságra vonatkozó valós adatait a modellek által generált virtuális mintákkal. Az elemzett tartalmi kérdés a demokráciával való elégedettség volt, amelyhez a legfontosabb magyarázóváltozók (a nem, az életkor, az iskolai végzettség, a jobb skálán való elhelyezkedés, illetve a jogrendszerbe vetett bizalom mértéke) alapján olyan virtuális válaszadókat generáltam, amelyek karakterjegyeikben meggyeztek a valós mintákban megfigyelt személyekkel, majd ezeket hasonlítottam össze egymással statisztikai eszközökkel.

Az eredmények alapján elmondható, hogy a nagy nyelvmodellek a ténylegeshez hasonló válaszokat generáltak, és előhívták a várt társadalmi mintázatokat a nekik megadott korlátozott kontextuális információból is, amennyiben számukra példákat biztosítottam, azonban példák nélkül nem teljesítettek elfogadhatóan. Az is láthatóvá vált, hogy nem mindegy, milyen bemenetekkel és beállításokkal használhatók ezek a modellek: a megadott példák száma és típusa (véletlen vagy a virtuális válaszadó szempontjából releváns példák) jelentősnek bizonyult a nagy nyelvmodellek teljesítményének sikerességében. Ezenkívül természetesen a modellek típusai is befolyásolták a teljesítményt, nem mindegyik modell teljesített egyformán jól. Jelen tanulmányban vizsgált modellek közül a GPT-4 mutatott jobb eredményeket, felülmúlva elődjét, a GPT-3.5-turbo-t. Az eredmények részletesebb vizsgálata azonban azt mutatta, hogy nem létezik egyértelműen legjobb modell. A legjobban teljesítő promptmodell kombinációnál találtam arra utaló jeleket, hogy a generált és a tényleges válaszok eltérése nem véletlenszerű. A torzítások jellegei a közismert szociológiai sztereotípiák szerint adódtak. Továbbá az is látható volt, hogy a különböző országok, illetve a célország nyelve szerint is van különbség a nagy nyelvmodellek teljesítményében.

A mesterséges intelligencia – többek között a nagy nyelvmodellek – által generált adatok használatával a korábbi kutatások ígérete szerint megoldást kínálhatunk a társadalomtudományos kutatások empirikus adatfelvételei körül kibontakozó aktuális problémákra, például az alacsony részvételi hajlandóság és válaszadási arány miatti tudományos érvényesség problémáira, tehát virtuálisan válaszadásra bírhatjuk azokat is, akiket a valóságban nem érünk el. A nagy nyelvmodellek által generált adatok észszerű és kritikus használatával hatékonyabbá tehetjük a társadalomtudományos kutatások előkészítési és tervezési fázisát (például virtuális pilotkutatásokkal), csökkenthetjük a valós adatfelvételekre fordított idő- és energiaigényt, továbbá

a hiányzó adatok pótlására (például a hiányzó értékek utólagos imputálásával) is megoldásokat dolgozhatunk ki.

Az általam végzett kísérletek eredményei ugyanakkor nem támasztják alá a túlzottan optimista, pozitívista elképzeléseket. A modellek példák nélkül is válaszoltak a kérdésekre, de ezen válaszok tudományos használhatósága kétséges. Példákkal kiegészítve jobb volt a helyzet, de ha ezekre szükség van, az jelentősen csökkenti az alkalmazások körét. A kérdőíves kutatások esetében leginkább a részleges válaszhiányok kezelése jöhet szóba, a szisztematikus torzítások forrásának alaposabb vizsgálata azonban ebben az esetben is szükséges (*Schramowski et al., 2022*). További kutatás tárgyát képezhetné, hogy amennyiben csak az aktívan internetezőkre szűkítjük a kiértékelést, akkor a valóshoz közelebbi válaszokat kapunk-e a modellekkel, így azonban le is mondanánk az internet nyelvi terében nem reprezentált társadalmi valóságról.

## Függelék

F1. tábla

### A különböző modellekkel és beállításokkal kapott KL-divergenciák közötti eltérések bootstrap $p$ -értékei a magyar mintára

*Bootstrap  $p$ -values of the differences between KL divergences obtained with different models and settings for the Hungarian sample*

	ESS magyar minta	GPT-4 5 célzott példa	GTP-4 5 véletlen példa	GPT-3.5 5 célzott példa	GPT-4 nincs példa	GPT-3.5 5 véletlen példa	GPT-3.5 nincs példa
ESS magyar minta	1	0	0	0	0	0	0
GPT-4 5 célzott példa	0	1	0,0065	0	0	0	0
GTP-4 5 véletlen példa	0	0,0065	1	0	0	0	0
GPT-3.5 5 célzott példa	0	0	0	1	0,335	0,0055	0
GPT-4 nincs példa	0	0	0	0,335	1	0,0065	0
GPT-3.5 5 véletlen példa	0	0	0	0,0055	0,0065	1	0,007
GPT-3.5 nincs példa	0	0	0	0	0	0,007	1

Forrás: az ESS-adatfelvétel magyar adatainak, illetve a kutatáshoz használt nagy nyelvmodellek által generált adatok statisztikai elemzése alapján saját elemzés és szerkesztés.

F2. tábla

### A különböző modellekkel és beállításokkal kapott KL-divergenciák közötti eltérések bootstrap $p$ -értékei a brit mintára

*Bootstrap  $p$ -values of the differences between KL divergences obtained with different models and settings for the British sample*

	ESS brit minta	GPT-4 5 célzott példa	GTP-4 5 véletlen példa	GPT-4 nincs példa	GPT-3.5 5 célzott példa	GPT-3.5 5 véletlen példa	GPT-3.5 nincs példa
ESS brit minta	1	0	0	0	0	0	0
GPT-4 5 célzott példa	0	1	0,015	0,0105	0	0	0
GTP-4 5 véletlen példa	0	0,015	1	0,4905	0,0085	0	0
GPT-4 nincs példa	0	0,0105	0,4905	1	0,0065	0	0
GPT-3.5 5 célzott példa	0	0	0,0085	0,0065	1	0,073	0,008
GPT-3.5 5 véletlen példa	0	0	0	0	0,073	1	0,2165
GPT-3.5 nincs példa	0	0	0	0	0,008	0,2165	1

Forrás: az ESS-adatfelvétel brit adatainak, illetve a kutatáshoz használt nagy nyelvmodellek által generált adatok statisztikai elemzése alapján saját elemzés és szerkesztés.

F3. tábla

**A GPT-4 modell válaszainak hibáját magyarázó általánosított  
additív modell kimenete a magyar és brit mintákra**  
*Output of the generalized additive model explaining  
the error of the GPT-4 model responses for the Hungarian and British samples*

ESS magyar minta				
	Becsült együttható	Standard hiba	<i>t</i> -érték	<i>p</i> -érték
Konstans	0,021	0,067	0,315	0,753
Nem (referencia: nő)	-0,006	0,032	-0,198	0,843
Életkor	-0,001	<0,001	-1,243	0,214
	Becsült szabadságfok	Referencia szabadságfok	F-statisztika értéke	F-statisztika <i>p</i> -értéke
Iskolai végzettség (spline)	1,0	1	0,562	0,454
Bal-jobb skálán mért beállítódás (spline)	8,3	9	2,899	0,001
Jogrendszerbe vetett bizalom (spline)	7,6	9	7,651	< 0,001
ESS brit minta				
	Becsült együttható	Standard hiba	<i>t</i> -érték	<i>p</i> -érték
Konstans	-0,229	0,104	-2,209	0,027
Nem (referencia: nő)	0,031	0,046	0,670	0,503
Életkor	0,001	0,001	1,103	0,270
	Becsült szabadságfok	Referencia szabadságfok	F-statisztika értéke	F-statisztika <i>p</i> -értéke
Iskolai végzettség (spline)	1,0	1	11,995	0,001
Bal-jobb skálán mért beállítódás (spline)	5,5	9	2,135	0,002
Jogrendszerbe vetett bizalom (spline)	8,4	9	11,543	< 0,001

Forrás: az ESS-adatfelvétel magyar és brit adatainak, illetve a kutatáshoz használt nagy nyelvmodellek által generált adatok statisztikai elemzése alapján saját elemzés és szerkesztés.

## Irodalom

- Aher, G. – Arriaga, R. I. – Kalai, A. T. (2023): *Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies*. arXiv:2208.10264.  
<https://doi.org/10.48550/arXiv.2208.10264>
- ANES (2021): *About us. About ANES*. American National Election Studies.  
<https://electionstudies.org/about-us/>
- Argyle, L. P. – Busby, E. C. – Fulda, N. – Gubler, J. – Rytting, C. – Wingate, D. (2022): Out of One, Many: Using Language Models to Simulate Human Samples. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 1(Long Papers): 819–862.  
<https://doi.org/10.18653/v1/2022.acl-long.60>

- 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>
- Brown, T. B. – Mann, B. – Ryder, N. – Subbiah, M. – Kaplan, J. – Dhariwal, P. – Neelakantan, A. – Shyam, P. – Sastry, G. – Askell, A. – Agarwal, S. – Herbert-Voss, A. – Krueger, G. – Henighan, T. – Child, R. – Ramesh, A. – Ziegler, D. M. – Wu, J. – Winter, C. – Amodei, D. (2020): *Language Models are Few-Shot Learners*. arXiv:2005.14165. <https://doi.org/10.48550/arXiv.2005.14165>
- ESS (2022): *About ESS – European Social Survey*. <https://www.europeansocialsurvey.org/about-ess>
- Garrido, A. (2009): About some properties of the Kullback-Leibler divergence. *Advanced Modeling and Optimization*, 11(4).
- Heersmink, R. – de Rooij, B. – Clavel Vázquez, M. J. – Colombo, M. (2024): A phenomenology and epistemology of large language models: Transparency, trust, and trustworthiness. *Ethics and Information Technology*, 26(3), 41. <https://doi.org/10.1007/s10676-024-09777-3>
- Mahowald, K. – Ivanova, A. A. – Blank, I. A. – Kanwisher, N. – Tenenbaum, J. B. – Fedorenko, E. (2024): Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6), 517–540. <https://doi.org/10.1016/j.tics.2024.01.011>
- Schramowski, P. – Turan, C. – Andersen, N. – Rothkopf, C. A. – Kersting, K. (2022): *Large Pre-trained Language Models Contain Human-like Biases of What is Right and Wrong to Do*. arXiv:2103.11790. <https://doi.org/10.48550/arXiv.2103.11790>
- Shanahan, M. (2023): *Talking About Large Language Models*. arXiv:2212.03551. <https://doi.org/10.48550/arXiv.2212.03551>
- Touvron, H. – Lavril, T. – Izacard, G. – Martinet, X. – Lachaux, M.-A. – Lacroix, T. – Rozière, B. – Goyal, N. – Hambro, E. – Azhar, F. – Rodriguez, A. – Joulin, A. – Grave, E. – Lample, G. (2023): *LLaMA: Open and Efficient Foundation Language Models*. arXiv:2302.13971. <https://doi.org/10.48550/arXiv.2302.13971>
- Wei, J. – Wang, X. – Schuurmans, D. – Bosma, M. – Ichter, B. – Xia, F. – Chi, E. – Le, Q. – Zhou, D. (2023): *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. arXiv:2201.11903. <https://doi.org/10.48550/arXiv.2201.11903>
- Yao, S. – Yu, D. – Zhao, J. – Shafran, I. – Griffiths, T. L. – Cao, Y. – Narasimhan, K. (2023): *Tree of Thoughts: Deliberate Problem Solving with Large Language Models*. arXiv:2305.10601. <https://doi.org/10.48550/arXiv.2305.10601>