

Sándor Anna Viola

Verbális agresszió értékelése mesterséges intelligenciával



Eötvös Loránd Tudományegyetem Pedagógiai és
Pszichológiai Kar

Témavezető: dr. Nagy Tamás

ELTE Stressz és Érzelmek Kutatócsoport

Absztrakt.....	3
Bevezető.....	4
Módszer.....	5
Résztevők.....	6
Eszközök.....	7
Kutatás folyamata.....	8
Agresszív válaszok gyűjtése.....	8
Emberi értékelések.....	9
Nagy nyelvi modellek értékelései.....	9
Adatfeldolgozás és statisztikai elemzés.....	10
Eredmények.....	12
Leíró statisztikák.....	12
Emberi értékelések.....	12
Nagy nyelvi modellek értékelései.....	14
Hipotézis ellenőrzés.....	16
Diskusszió.....	21
Eredmények összefoglalása.....	21
Vizsgált hipotézisek.....	21
Hiányzó értékelések.....	22
Költségek.....	22
Erősségek és limitációk.....	23
Konklúzió.....	24
Záró megjegyzések.....	24
Irodalomjegyzék.....	24

Absztrakt

A verbális agresszió egy digitális terekben különösen gyakori jelenség, amely számos pszichológiai problémához vezethet. Ennek megfelelően detektálása az utóbbi időben egyre fontosabb lett. Kutatásom célja annak vizsgálata volt, hogy az ismert és széles körben használt nagy nyelvi modellek (ChatGPT-4o, ChatGPT-4o mini, Claude 3.5 Sonnet, Claude 3.5 Haiku, Gemini 1.5 Flash, Gemini 1.5 Pro, Llama 3.2 3B, Llama 405B) képesek-e az emberekhez hasonlóan értékelni mondatok verbális agresszivitását. Azt feltételeztük, hogy az emberi és a nagy nyelvi modellek értékelései között 0,6 feletti erős korreláció lesz (H1), továbbá nem lesz szignifikáns különbség az emberi és nagy nyelvi modellek értékeléseinek mondatonkénti átlaga (H2) és szórása (H3) között.

A hipotéziseket egy három részből álló kutatásban vizsgáltuk, melynek első fázisában potenciálisan agresszív mondatokat gyűjtöttünk személyes részvétellel, majd ezeket értékeltettük egy online kérdőívben emberekkel, illetve APA híváson keresztül a vizsgált nagy nyelvi modellekkel. A hipotéziseket Pearson-korrelációval (H1) és generalizált kevert-hatású gamma regresszióval teszteltük (H2, H3).

A Llama 3.2 3B modell kivételével minden modell esetén erős, 0,6 feletti korrelációt találtunk az emberi értékelésekkel, a ChatGPT-4o mini ($\rho=0,81$), a Claude 3.5 Sonnet ($\rho=0,82$) és a Gemini 1.5 pro ($\rho=0,83$) esetén pedig a korreláció 0,8 feletti, nagyon erős volt. Az átlagra és szórásra vonatkozó hipotéziseket azonban csak korlátozottan sikerült igazolnunk. Minden vizsgált modell értékeléseinek mondatonkénti átlaga szignifikánsan különbözött az emberi értékelésektől, míg a szórások esetén csak a Llama 3.2 3B modell esetén nem volt szignifikáns a különbség ($CI=0,91-0,94$). A modellek ennek ellenére a magas korrelációs értékekből látható módon konzisztensen ítélték meg a vizsgált mondatokat, emiatt jól használhatóak lehetnek a verbális agresszió detektálására.

A kutatás az ELTE PPK Stressz és Érzelmek Kutatócsoportjának munkájaként valósult meg. Személyesen a kutatásban annak megtervezésétől kezdve közreműködtem. Részt vettem az agresszív mondatokat eredményező személyesen zajló adatgyűjtésben, én készítettem az emberi agresszióértékeléseket összegyűjtő kérdőívet, illetve írtam meg a modellek értékeléséhez szükséges Python kódot. Továbbá részt vettem az adatok elemzésében, a következtetések megfogalmazásában, illetve én írtam meg a dolgozatot.

Bevezető

Verbális agressziónak azt a kommunikációs formát nevezzük, mely során valaki expliciten megtámadja egy másik személy önképét (Aloia és mtsai., 2020), és melynek célja pszichológiai fájdalom okozása (Heisel, 2010). A verbális agresszió korunk egy fontos problémája, ami a digitális médiában különösen gyakorivá vált (Appel és mtsai., 2014; Bykov és mtsai., 2018). A verbális agresszió számos pszichológiai problémához, szorongáshoz, illetve depresszióhoz vezethet (Cole és mtsai., 2014; Hanley & Gibb, 2011; Rueger & Jenkins, 2014), továbbá gyakran összekapcsolódik fizikai agresszióval is (Poling és mtsai., 2019).

Ennek megfelelően a verbális agresszió detektálása egyre fontosabbá vált a digitális és valódi terekben egyaránt (lásd például: Hengel & Andringa, 2007). Számos tanulmány vizsgálta, hogy gépi tanulással (összefoglalásként lásd: Del Bosque & Garza, 2014), illetve deep learning módszerekkel eredményesen lehet-e detektálni az online felületeken és közösségi médiában megjelenő agressziót (Sadiq és mtsai., 2021), és az utóbbi módszerrel több esetben 80-90%-os pontosságot értek el (Sadiq és mtsai., 2021). Ennek ellenére jelenleg még nem áll rendelkezésünkre olyan modell, amelyet széles körben használnának a verbális agresszió detektálására.

Kutatásom célja annak vizsgálata, hogy az ismert és széles körben használt nagy nyelvi modellek képesek lehetnek-e eredményesen detektálni és értékelni mondatok verbális agresszivitását. Ezzel kapcsolatban még viszonylag kevés tanulmány született. A legismertebb mesterséges intelligencia modell, a ChatGPT esetén Kocoń és mtsai. (2023) tanulmánya szerint 77.91% pontossággal tudta eldönteni egy mondatról, hogy agresszív-e. Más tanulmányok a gyűlöletbeszéd (Wen & Younes, 2024) és a bántó nyelvezet (Bauer és mtsai., 2024) detektálásában értek el sikereket a ChatGPT használatával. Nina-Gutiérrez és mtsai. (2024) a ChatGPT és a Google Gemini modelljével is hasonló eredményt értek el a verbális agresszió felismerésében. Trillo és mtsai. (2024) a Gemini modellt a verbális agresszió detektálásán keresztül eredményesen használták csoporton belüli konfliktusok csökkentésére. A Meta nagy nyelvi modellje, a Llama olasz verziója pedig eredményesen ismerte fel a párkapcsolati erőszakot (Basile és mtsai., 2024).

A biztató eredmények ellenére eddig nem született olyan átfogó kutatás, amely az említett modellek képességét vizsgálná a verbális agresszió értékelésére. Kutatásom célja ennek megfelelően annak vizsgálata volt, hogy az ismert és széles körben használt nagy nyelvi modellek agresszió értékelése hasonlít-e a humán értékelők által adott agresszió

pontszámokra, ezáltal kiválthatja-e azokat. Amennyiben igen, a modellek használata jelentősen megkönnyíthetné a verbális agresszió vizsgálatával foglalkozó kutatók munkáját, illetve fine-tuning módszerekkel lehetővé tenné egy széles körben használható agressziófelismerő és értékelő modell létrehozását.

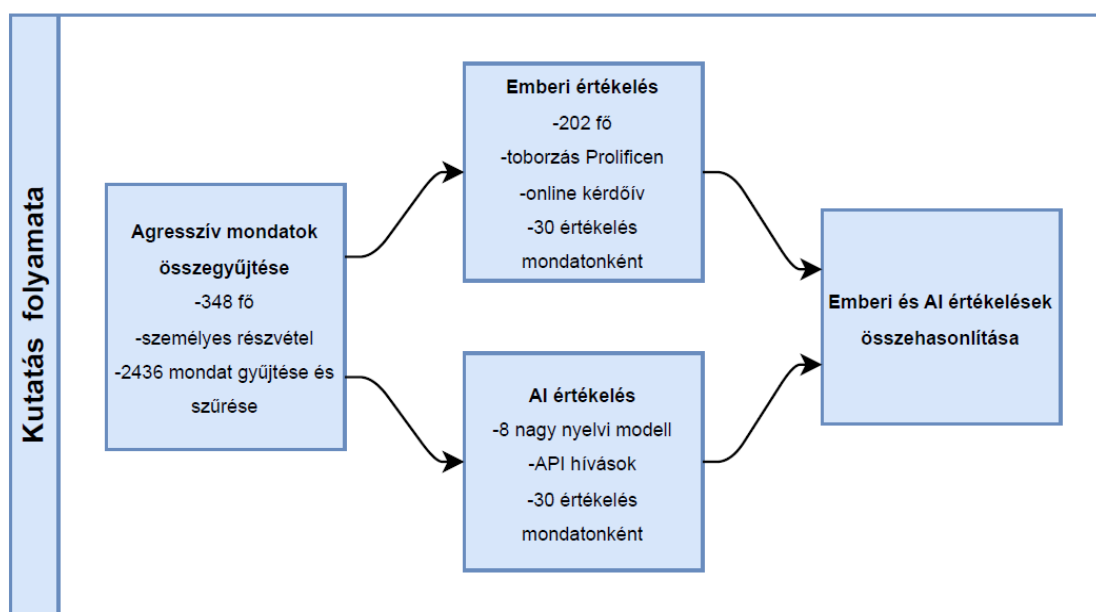
Kutatásomban az OpenAI ChatGPT-4o, illetve ChatGPT-4o mini modelljét (OpenAI, 2024, <https://chatgpt.com/>), a Google Gemini 1.5 Pro és 1.5 Flash modelljét (Google, 2024, <https://gemini.google.com/app>), az Anthropic Claude 3.5 Sonnet és 3.5 Haiku modelljét (Anthropic, 2024, <https://claude.ai/new>) és a Meta Llama 405B és a Llama 3.2 3B modelljét (Llama, 2024, <https://www.llama.com/>) vizsgáltam. A modelleket az alapján választottuk ki, hogy a kutatás preregisztrálásakor (Open Science Framework, Charlottesville: <https://osf.io/tpzw9>) a legismertebb nagy nyelvi modellek elérhető legnagyobb teljesítményű modelljét, illetve azok legjobb teljesítményű kis verzióját is vizsgáljuk.

A hipotéziseink ezek alapján a vizsgált modellekre a következők voltak. Feltételeztük, hogy pozitív, erős (0.6 feletti) korreláció van az emberi és a nagy nyelvi modellek által adott értékelések között (H1), továbbá feltételeztük, hogy nem lesz szignifikáns különbség a humán és a nagy nyelvi modellek értékeléseinek átlaga (H2), illetve szórása között (H3).

Módszer

A kutatás három részből állt. Az első részben potenciálisan agresszív válaszokat gyűjtöttünk, majd ezeket a kutatás második fázisában az első vizsgálatától független résztvevőkkel értékeltettük egy online kérdőívben. Végül ugyanezeket a válaszokat a vizsgált nagy nyelvi modellekkel is értékeltettük, és az így kapott értékeléseket összehasonlítottuk az emberi értékelésekkel.

1. Ábra: A kutatás folyamata



Résztvevők

Az agresszív válaszok gyűjtését célzó kísérletben 348 fő vett részt. A minta elemszáma a kutatásban szereplő másik hipotézishez, (mely nem témája jelen dolgozatomnak) igazodott. A kutatáshoz kényelmi mintavételt alkalmaztuk: a minta az ELTE hallgatóiból állt, akik a részvételükért részleges kurzus kreditet kaptak. A kutatásban 88 férfi és 260 nő vett részt. Életkoruk 18 és 48 év között volt (átlag: 21, szórás: 3.8).

A kutatás második, emberi értékelés fázisában 202 fő vett részt. A résztvevők számát úgy határoztuk meg, hogy mondatonként 30 értékelést tudjunk összegyűjteni, miközben minden résztvevő képenként 50 mondatot értékel (ennek megfelelően összesen 350 mondatot). Az egy fő által értékelendő mondatok számát az előzetes tesztelések során úgy határoztuk meg, hogy a kérdőív elég rövid maradjon ahhoz, hogy a résztvevők kifáradása ne befolyásolja az értékeléseket. A résztvevőket a Prolific (Prolific, London: www.prolific.com) felületén toboroztuk, ahol a kitöltésért pénzt kaptak. A résztvevők között 101 nő és 100 férfi volt, 1 résztvevő pedig nem nyilatkozott a neméről. Életkoruk 18-81 év volt (átlag: 41.5 év, szórás: 14.5).

A két emberi résztvevővel zajló adatgyűjtéshez az etikai engedélyeket az ELTE PPK Kutatásetikai Bizottsága adta ki.

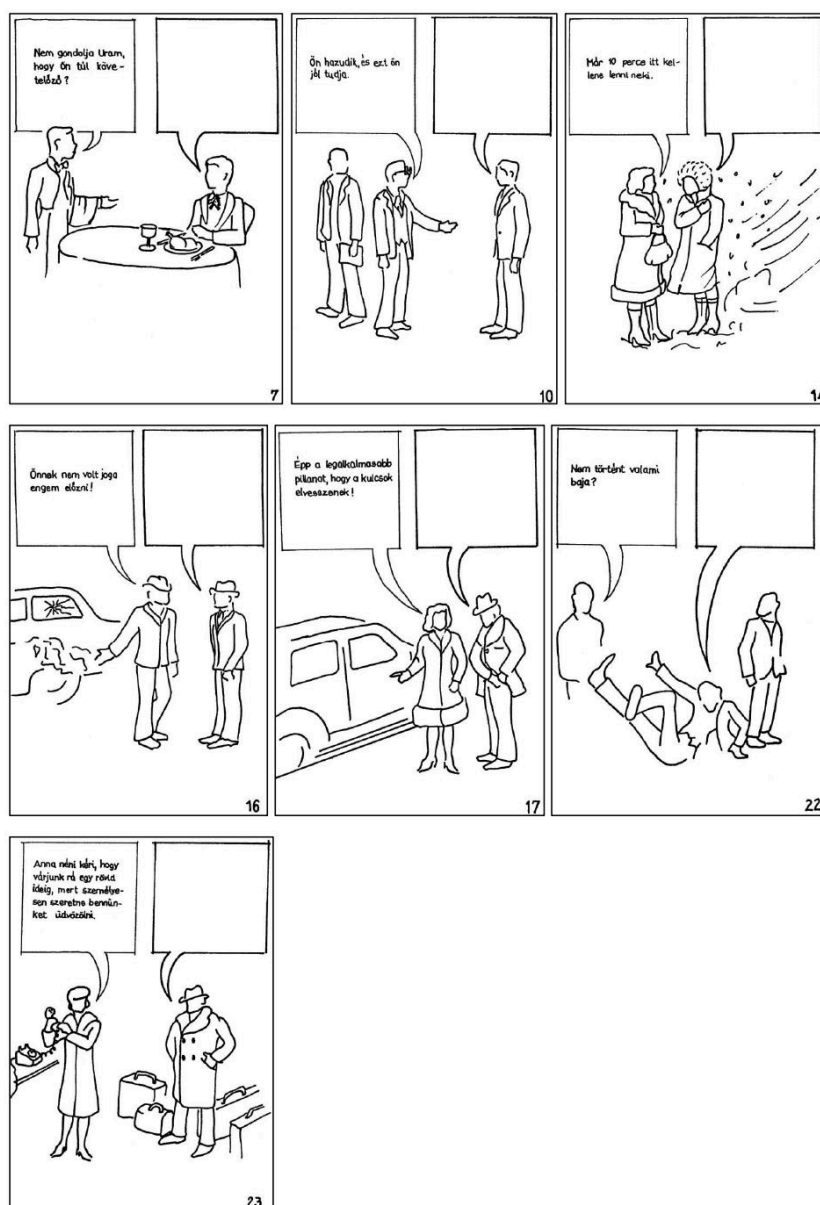
Eszközök

A verbálisan agresszív válaszok indukálására a Rosenzweig-féle képfrusztrációs teszt hét itemét (7., 10., 14., 16., 17., 22., 23.) használtuk (Rosenzweig és mtsai., 1975). A teszt lényege, hogy a résztvevők egy olyan provokatív szituációt ábrázoló képpel szembesülnek (pl. elrontott rendelést hoz ki a pincér, összekoccan két autó), amelyben a képen szereplő egyik személy egy provokatív megjegyzést tesz. A résztvevők feladata, hogy a másik személy helyébe képzelve magukat megfogalmazzák a válaszukat az adott szituációban. A teszt indukálhat agresszív és kevésbé agresszív válaszokat is. Az általunk használt itemek a 2. ábrán láthatók.

A válaszok agresszivitásának emberi értékeltetése egy Qualtrics felületén futó kérdőív segítségével történt (Qualtrics, Provo UT: <https://www.qualtrics.com>), melyhez a résztvevőket a Prolific felületén (Prolific, London: www.prolific.com) keresztül toboroztunk. A válaszok angolra fordításához pedig gépi fordító segítségét is igénybe vettük (DeepL Translate, Köln: <https://www.deepl.com/en/translator>).

A nagy nyelvi modellek agresszióértékeléseinek összegyűjtése Python kódon keresztül futtatott API hívások segítségével zajlott az adott nagy nyelvi modellt forgalmazó cégtől igényelt API kulcson keresztül (kivéve a Llama 405B esetén, melyhez a FireworksAI oldalt használtuk - FireworksAI, Redwood City: <http://fireworks.ai>). Az értékelések összehasonlítását célzó statisztikai elemzést R-ben végeztem el (R: The R Project for Statistical Computing, Bécs: <https://www.r-project.org/>).

2. Ábra: A Picture Frustration Test általunk használt itemei



Kutatás folyamata

Agresszív válaszok gyűjtése

Az agresszív mondatok gyűjtése kísérleti elrendezésben zajlott. A vizsgálat jelentős része a jelen dolgozat témájától független kérdésfeltevések vizsgálatára irányult. A kísérlet során a résztvevők más kérdőívek mellett kitöltöttek egy félprojektív tesztet, melyben írásbeli válaszokat kellett adniuk hét különböző kétértelmű szituációra (Rosenzweig-féle Picture

Frustration Test, Rosenzweig és mtsai., 1975). Az így összegyűjtött mondatok szolgáltatták a jelen vizsgálat kiindulópontját.

A kutatás első fázisa 348 fő részvételével zajlott, mely során összesen 2436 szöveges választ gyűjtöttünk. Ezek közül 32 válasz az értékelést megelőzően kizárásra került. A kizárás okai a következők voltak: a résztvevők egy része nem jól értelmezte a feladatot (például a saját érzéseiről írt a szituációba helyezkedés helyett), vagy válaszuk egyáltalán nem kapcsolódott a szituációhoz, esetleg annyira helytelen volt nyelvtanilag, hogy nem lehetett értelmezni.

Emberi értékelések

A kutatás második fázisában a megmaradó 2404 szöveges választ értékeltettük emberi résztvevőkkel egy online kérdőívben. A magyar nyelvű válaszokat az értékelést megelőzően angol nyelvre fordítottunk, hogy az összes nagy nyelvi modell értékelni tudja őket. A mondatokat a DeepL gépi fordító segítségével (DeepL Translate, Köln: <https://www.deepl.com/en/translator>) fordítottuk, majd a fordítást a kutatócsoport három tagja ellenőrizte, és hagyta jóvá. A fordítás miatt az értékelő kérdőív kitöltésének feltétele az angol nyelvtudás volt.

A kérdőív a következőképpen épült fel. A résztvevők az oldal tetején látták a képet, melyre az adott válaszok születtek, majd képenként 50 választ kellett értékelniük egy 0-tól 10-ig terjedő skálán az alapján, hogy az adott mondatot mennyire találták agresszívnek (0=egyáltalán nem agresszív, 10=nagyon agresszív). Egy résztvevő a hét képnek megfelelően összesen 350 véletlenszerűen kiosztott szöveges választ értékelt.

A kérdőívet úgy állítottuk be, hogy minden választ körülbelül 30 különböző személy értékelje (az egy képhez tartozó mondatok eltérő száma miatt a kapott emberi értékelések száma végül 26 és 36 között változott). Az így kapott értékeléseket hasonlítottuk össze a kutatás további részében a nagy nyelvi modellek ugyanarra a mondatra adott értékeléseivel.

Nagy nyelvi modellek értékelései

A kutatás harmadik fázisában az összegyűjtött válaszokat értékeltettük a vizsgált nagy nyelvi modellekkel. A modellek többségét publikus API hívással értük el, melyhez egy Python kódot használtunk. A Meta Llama open source 3B modellje saját gépen futott, míg a 405B szintén open source modellt a személyi számítógépeket meghaladó teljesítményigénye miatt a FireworksAI oldalon keresztül futtattuk (FireworksAI, Redwood City: <http://fireworks.ai>).

A nagy nyelvi modelleknek adott instrukciót úgy fogalmaztuk meg, hogy a legjobban hasonlítson az emberi értékelőknek adott instrukciókra. A modelleket arra instruáltuk, hogy egy motivált emberi válaszadóhoz hasonlóan értékeljék a megadott mondatok agresszivitását 0 és 10 között (0=egyáltalán nem agresszív, 10=nagyon agresszív) az adott szituáció figyelembe vételével, (melyet a csatolt kép jelölt), továbbá megadtuk, hogy a válaszuk egész szám formátumú legyen, és a költségek csökkentése érdekében a válasz maximum két tokenből álljon. Az instrukcióhoz (prompt) az emberi értékelítéshez hasonlóan az adott válaszhoz tartozó képet is csatoltuk. A Llama modellek esetén a képfeldolgozás nem elérhető Európán belül, így a 3B és a 405B modell esetén a képekre vonatkozó részt töröltük az instrukcióból.

Az instrukció ez alapján a következő volt a képeket értelmezni képes modellek esetén: *“Respond as a motivated human rater who rates verbal aggression. You must respond with ONLY a number from 0 to 10 in integer format, nothing else. You can see a picture of a situation, where a person says something to another person. The second person answers (blank text bubble). You will also see the answer belonging to the situation. Your task is to rate this sentence's aggressiveness from 0 to 10 (0 = not aggressive, 10 = extremely aggressive): Respond with ONLY the number in integer format, nothing else.”*, illetve kép nélkül: *“Respond as a motivated human rater who rates verbal aggression. Respond with ONLY a number from 0 to 10 in integer format, nothing else. Your task is to rate this sentence's aggressiveness from 0 to 10 (0 = not aggressive, 10 = extremely aggressive)”*.

A mondatokat egyesével értékeltettük a nagy nyelvi modellekkel, ezzel gátolva, hogy a korábbi válaszok befolyásolják a későbbieket. A modellek temperature értékét az alap (default) értéken hagytuk, hogy ezzel se befolyásoljuk, optimalizáljuk az eredményt. Továbbá, annak érdekében, hogy az emberi értékelésekkel legjobban összehasonlítható adatokat kapjunk, minden mondatot 30-szor értékeltettünk le az összes modellel, (a modellek körönként leértékelték mind a 2404 választ, majd újrakezdték az értékelést). Az így kapott pontszámokat hasonlítottuk össze a korábban összegyűjtött emberi értékelésekkel.

Adatfeldolgozás és statisztikai elemzés

Az adatfeldolgozáshoz az R programnyelvet használtuk (*R Studio*, Boston: <http://www.rstudio.com/>). Az elemzéshez használt csomagok a következők voltak: tidyverse (Wickham & RStudio, 2023), fs (Hester és mtsai., 2024), skimr (Waring és mtsai., 2022), glmmTMB (Brooks és mtsai., 2024), correlation (Makowski és mtsai., 2024), performance

(Lüdecke, Makowski, és mtsai., 2024), sjPlot (Lüdecke, Bartel, és mtsai., 2024), broom (Robinson és mtsai., 2024).

Az adatfeldolgozás első lépéseként összekapcsoltuk az emberi és a nagy nyelvi modellek által adott értékeléseket, majd a továbbiakban ezzel az adattáblával dolgoztunk. Az emberi értékelések közül a kérdőív formátumából kiindulva (az értékeléseket a felajánlott lehetőségek közül választották ki a 0-tól 10-ig terjedő skálán) egyet sem kellett kizárnunk. A nagy nyelvi modellek esetén a kizárás kritériumát a preregisztrációban úgy határoztuk meg, hogy azokat az értékeléseket zárjuk ki, melyek nem egész szám formátumúak, nem a megadott skálán helyezkednek el, illetve nem egy megadott mondathoz kapcsolódnak. Az ezek alapján kizárt értékek az eredmények részben láthatók.

A korrelációs hipotézishez (H1) Spearman-korrelációt használtunk, mert az emberi értékelések elemzése során kiderült, hogy az értékelések eloszlása nem normális, hanem jobbra ferde eloszlást mutat. Ezt a továbbiakban a dolgozat leíró statisztikák részében ismertetem.

Az átlagok (H2) és szórások különbségére (H3) vonatkozó hipotéziseket egy generalizált kevert-hatású gamma regresszióval teszteltük logaritmikus kapcsolati függvénnyel, amely az átlagra és a szóródásra (dispersion) vonatkozó hipotéziseket is teszteli egyszerre (továbbiakban: location-scale modell). A modell kimeneti változója az értékelés volt (a 0-tól 10-ig terjedő skálán), a prediktor pedig a modell típusa (emberi vagy nagy nyelvi modellek egyike). Mivel a random intercept és slope modell nem konvergált, a preregisztrációban rögzítettek szerint random intercept modellt használtunk. Ebben az értékelő azonosítója és a szituációba ágyazott válaszok voltak a csoportosító változók. Ez lehetővé tette, hogy a modell saját interceptet számoljon minden értékelőre, szituációra és válaszra. Erre a megoldásra a kimeneti változó jobbra ferde, korlátos és diszkrét eloszlása miatt volt szükség.

A preregisztrációban is rögzített módon (<https://osf.io/tpzw9>) az eredmények értelmezését a következőképpen határoztuk meg. A korrelációs hipotézishez (H1) a Spearman-féle korrelációs együttható becslése körüli kétfarkú 90%-os konfidenciaintervallumokat számoltunk. Hipotézisünk akkor igazolódott, ha a konfidenciaintervallum alsó határa nagyobb volt, mint 0,6. Ha a konfidenciaintervallum felső határa kisebb volt, mint 0,6, a hipotézisünket elutasítottuk. Ellenkező esetben a teszt eredményét az általunk meghatározott kritériumok alapján nem eldönthetőnek tekintettük.

Az emberi és nagy nyelvi modellek értékelései átlagának (H2) és szórásának (H3) különbségeivel kapcsolatos hipotézisek esetében az egyes értékelőtípusokhoz (ember vagy

nagy nyelvi modellek egyike) kapcsolódó hatásméretekre kétoldali 95%-os konfidenciaintervallumokat becsültünk. A statisztikai modellben az emberi értékelők szolgáltak referenciaszintként, a nagy nyelvi modellek együttthatói pedig exponenciális transzformáció után relatív arányoknak (RR) feleltek meg. Mivel az általunk választott statisztikai modell multiplikatív skálán működött, a következtetések egy előre meghatározott legkisebb érdekes hatásméretre (SESOI) támaszkodtak, amely a legkisebb értelmes eltérést jelölte a különbség hiányától. Ezt a preregisztrációban 0,5 pont abszolút különbségként határoztuk meg. A SESOI az emberi értékelések átlaga és a 0,5 egységnyi abszolút különbség arányán alapult, ami 0,864 és 1,136 SESOI küszöbértéket eredményezett egy multiplikatív skálán (azaz $RR = 1 \pm$ körülbelül 13,6%). Ezek a küszöbértékek jelentették a kutatásunk szempontjából legkisebb értelmes arányos változást a referencia (emberi) értékelésekhez képest.

Az átlagra és szórásra vonatkozó hipotéziseinket ezek alapján a következőképpen értékeltük. Akkor tekintettük úgy, hogy a vizsgált nagy nyelvi modell és az emberi értékelés között nincs különbség, ha az RR-bebecslés 95%-os konfidencia intervalluma teljes mértékben a [0,864, 1,136] tartományba esett. Ez felelt meg az általunk választott 0-tól 10-ig terjedő skálán a maximum 0,5 pontnyi különbségnek. A nagy nyelvi modellek és az emberi értékelések közötti különbséget abban az esetben tekintettük szignifikánsnak, ha a 95%-os konfidencia intervallum nem tartalmazta az 1-et. Abban az esetben, ha a 95%-os konfidencia intervallum a SESOI-tartományt ([0,864, 1,136]) és az 1-et is tartalmazta, az eredményt nem eldönthetőnek tekintettük. A fentebb ismertetett interferencia kritériumok az elemzés előtt a kutatás preregisztrációjában rögzítésre kerültek.

Eredmények

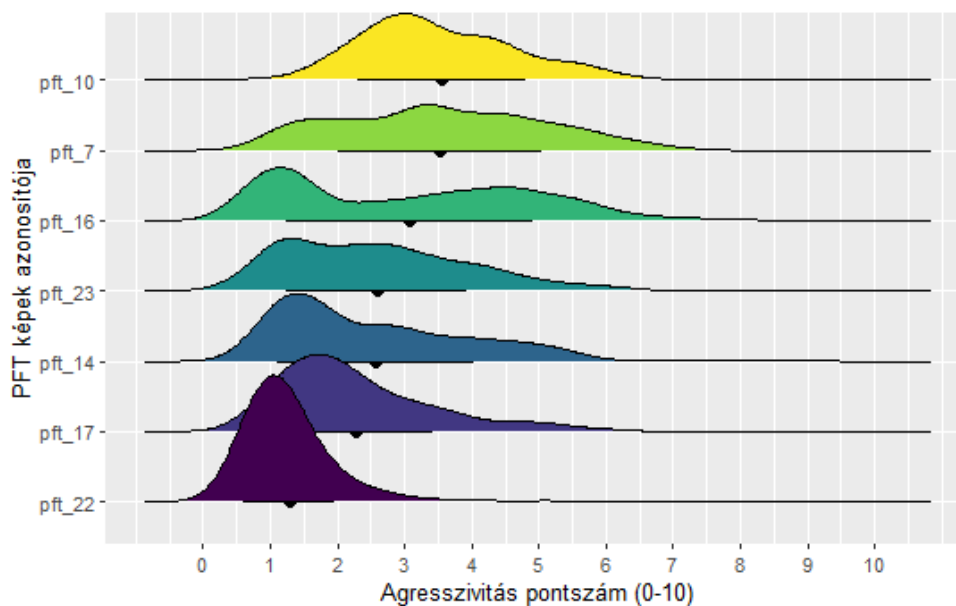
Leíró statisztikák

Emberi értékelések

A kérdőívben összesen 2404 választ értékeltettünk. A cél az volt, hogy válaszonként körülbelül 30 értékelést kapjunk. Egy mondatra érkezett legkevesebb értékelés 26, a legtöbb 36, a medián pedig 31 volt (a különbség a Qualtrics véletlen hozzárendelési algoritmusából eredt).

A mondatok értékelése a következőképpen alakult: 0-tól 10-ig terjedő skálán a 7-es kép átlag agressziópontszáma 3,46 (sd=1,41), a 10-es kép pontszáma 3,42 (sd=1,21), a 14-es képé 2,53 (sd=1,48), a 16-os képé 3,07-es (sd=1,83), a 17-es képé 2,3 (sd=1,15) a 22-es képé 1,25 (sd=0,661), a 23-as kép agressziópontszáma pedig 2,42 volt (sd=1,31), lásd 3. ábra.

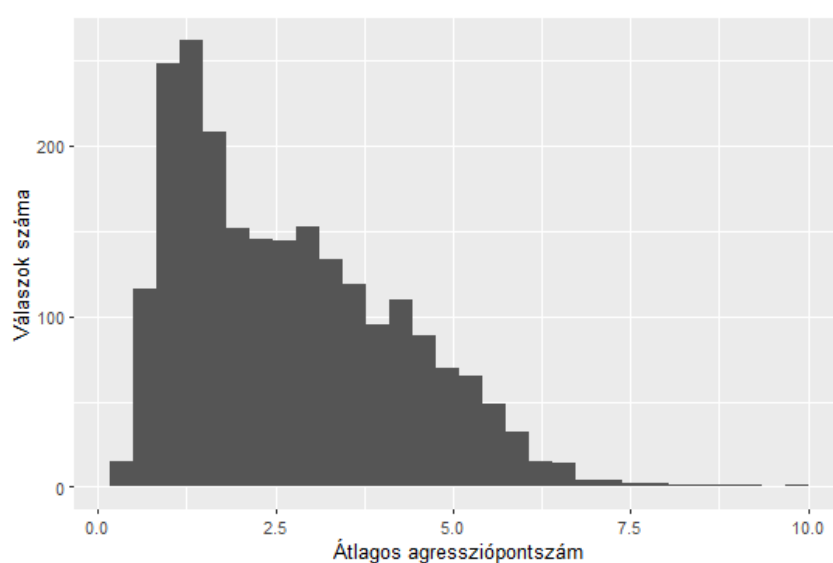
3. Ábra: A Picture Frustration Teszt képeire adott válaszok agressziópontszámának átlaga, eloszlása és szórása



Megjegyzés: A fekete pontok az átlagot, a vonalak a szórást jelzik.

Az átlagosan legagresszívebb válaszok a 7-es és a 10-es képre születtek, míg a legkevésbé agresszívek a 22-esre. Az átlag agressziópontszám azonban minden esetben alacsonyabb volt, mint a skála közepe. Ennek ellenére azonban született néhány magas agressziópontszámot kapott válasz is, amely az alábbi ábrán is jól látszik. Az összes válasz közül a legalacsonyabb elért átlagos agressziópontszám 0,22, a legmagasabb pedig 9,74 volt (a medián 2,38).

4. ábra: Válaszok átlagos agresszivitása az emberi értékelések alapján



Nagy nyelvi modellek értékelései

Az összegyűjtött mondatokat 8 nagy nyelvi modellel értékeltettük. Az összegyűjtött értékelések jellemzői az 1. táblázatban láthatók.

1. táblázat: Nagy nyelvi modellek értékeléseinek leíró statisztikája összehasonlítva az emberi értékelésekkel

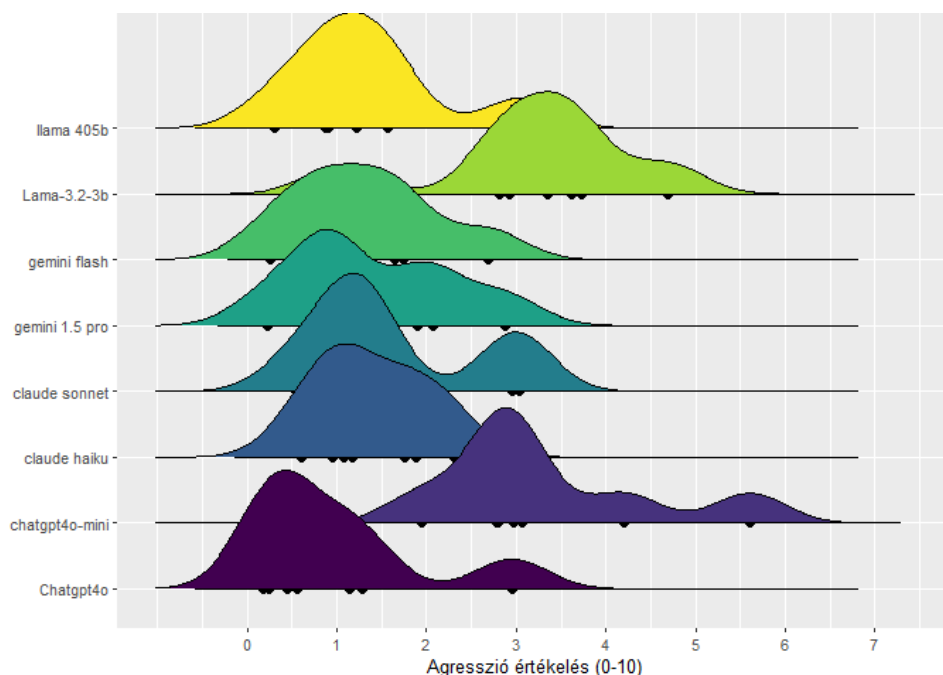
Típus	Kitöltöttség	Átlag	Szórás	p0	p25	p50	p75	p100
Emberi	100%	2.70	2.68	0	0	2	5	10
ChatGPT-4o	100%	0.95	1.54	0	0	0	2	10
ChatGPT-4o mini	98%	3.30	2.11	0	2	3	5	10
Claude 3.5 Haiku	100%	1.39	1.41	0	0	1	2	10
Claude 3.5 Sonnet	96%	1.61	1.95	0	0	1	3	10
Gemini 1.5 Pro	100%	1.39	1.69	0	0	1	2	10
Gemini Flash	100%	1.30	1.42	0	0	1	2	10
Llama 3.2 3B	100%	3.18	2.49	0	2	2	6	10

Llama 405B	99%	1.33	1.86	0	0	0	2	10
------------	-----	------	------	---	---	---	---	----

Megjegyzés: A p a pertcentilist jelöli. A kitöltöttség a meglévő értékelések arányát mutatja az összes kért értékeléshez képest (összesen 72 120). ChatGPT-4o mini esetén 1393, Claude Sonnet esetén 2465 és a Llama 405B esetén 943 érték hiányzott, melynek oka a módszer hibájából adódó adatvesztés (ChatGPT) vagy a modell korlátai voltak (részletesen lásd: Diskusszió).

A nagy nyelvi modellek által adott értékelések átlagát és szórását az 5. ábra is szemlélteti. A nagy nyelvi modellek értékelései az emberekhez hasonlóan jobbra ferde eloszlást mutattak, és a legtöbb értékelés 0 és 2 között változott. A legalacsonyabb pontszám minden modell esetén a minimális 0 volt, míg a legmagasabb a maximális 10. A legalacsonyabb átlagos agressziópontszámot a ChatGPT-4o modell esetén kaptuk (0,95), míg a legmagasabbat a ChatGPT-4o mini esetén (3,3). Az átlagos értékelés csak egy modell egy iteme (ChatGPT-4o mini) esetén haladta meg a skála közepét jelentő 5-t. Az értékelések szórása 1,41 (Claude 3.5 Haiku) és 2,41 (Llama 3.2 3B) között mozogtak, (az emberi értékelések szórása összehasonlításként 2,68 volt).

5. ábra: Nagy nyelvi modellek által adott értékelések átlaga, eloszlása és szórása



Megjegyzés: Az ábrán a fekete pontok a különböző képekre adott mondatok értékeléseinek átlagát jelölik

Hipotézis ellenőrzés

Az első hipotézisünket, mely szerint erős, 0.6 feletti korreláció van az emberi és a nagy nyelvi modellek által adott értékelések között, a preregisztrációban meghatározott módon Spearman korrelációval teszteltük. A korrelációs teszt eredményei az alábbi táblázatban láthatók (2. táblázat).

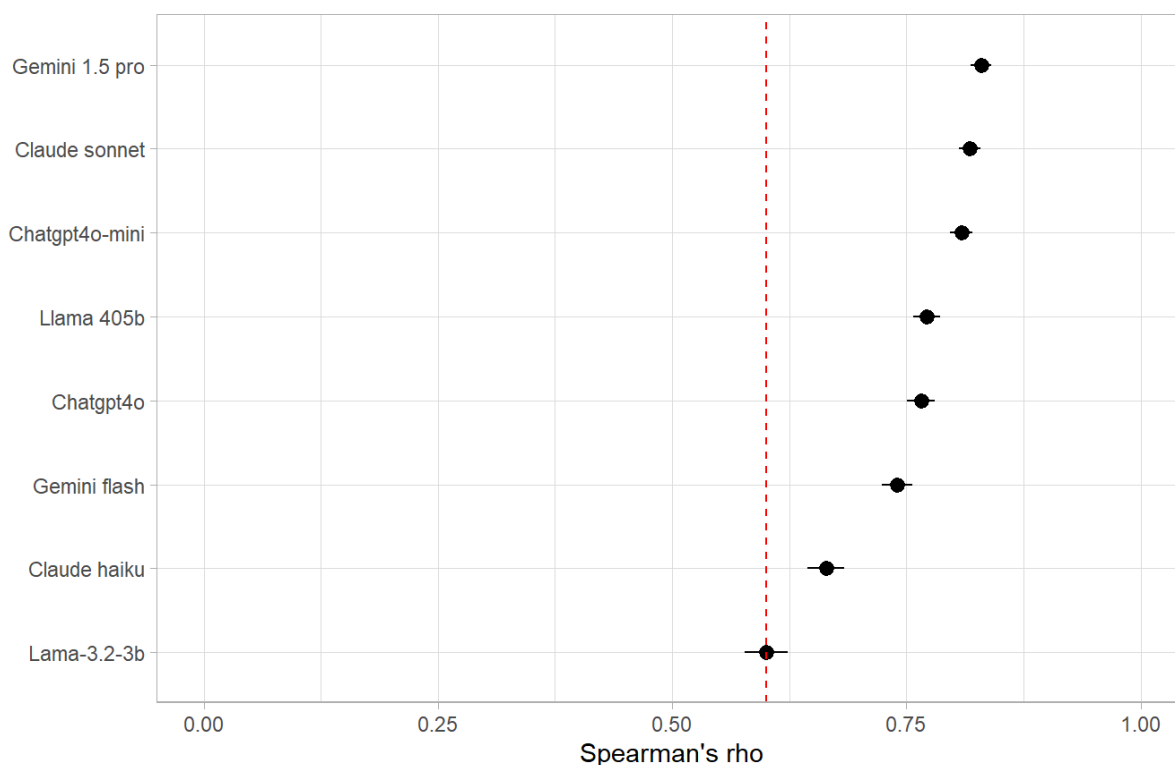
2. táblázat: Spearman-korreláció eredményei az emberi értékelésekkel a különböző modellekre

Modell neve	rho	CI_alsó	CI_felső	S	p-érték	n
ChatGPT-4o	0.77	0.75	0.78	445541018	<0.001	2251
ChatGPT-4o mini	0.81	0.80	0.82	363454268	<0.001	2251
Claude 3.5 Haiku	0.66	0.64	0.68	637491956	<0.001	2251
Claude 3.5 Sonnet	0.82	0.81	0.83	345421630	<0.001	2250
Gemini 1.5 Pro	0.83	0.82	0.84	323530376	<0.001	2251
Gemini Flash	0.74	0.72	0.76	493807049	<0.001	2251
Llama 3.2 3B	0.60	0.58	0.62	759091847	<0.001	2251
Llama 405B	0.77	0.76	0.79	432945793	<0.001	2250

Megjegyzés: 90%-os konfidencia intervallummal számolva

A hipotézist a Llama 3.2 3B modell kivételével ($\rho=0,6$) minden modell esetében igazoltuk (lásd 6. ábra). A Claude 3.5 Haiku ($\rho=0,66$), a Gemini Flash ($\rho=0,74$), a ChatGPT-4o ($\rho=0,77$) és a Llama 405B ($\rho=0,77$) esetén erős, 0,6 feletti korrelációt találtunk, míg a ChatGPT-4o mini ($\rho=0,81$), a Claude 3.5 Sonnet ($\rho=0,82$) és a Gemini 1.5 Pro ($\rho=0,83$) esetén nagyon erős, 0,8 feletti korrelációt találtunk az emberi értékelésekhez képest. A legalacsonyabb korreláció a Llama 3.2 3B modell esetén volt ($\rho=0,6$), míg a legmagasabb korreláció a Gemini 1.5 pro esetén ($\rho=0,83$). Általánosan elmondható, hogy az adott cég modelljei közül a kisebb modell gyengébb korrelációt mutatott az emberi értékelésekkel (kivéve az OpenAI ChatGPT modelljei esetén).

6. ábra: Spearman korreláció eredménye



Megjegyzés: A kritériumként meghatározott értéket a piros vonal jelöli, a 90%-os konfidencia intervallumokat pedig a hibásávok (error barok). A hipotézist a Llama 3.2 3B modell kivételével minden modellre igazoltuk.

A második hipotézisünket, mely szerint nincs szignifikáns különbség az emberi értékelések és a nagy nyelvi modellek értékelésének mondatonkénti átlaga között, a preregisztrációban meghatározott módon egy location-scale modellel teszteltük. A modell jellemzői az alábbi táblázatban láthatók.

3. Táblázat: A modell átlagra vonatkozó jellemzői

Prediktorok	Regressziós együtthatók	CI	Statisztika	p-érték
(Intercept)	3.20	2.56 – 4.00	10.24	<0.001
ChatGPT-4o	0.53	0.47 – 0.59	-11.23	<0.001
ChatGPT-4o mini	1.24	1.11 – 1.39	3.81	<0.001
Claude 3.5 Haiku	0.69	0.61 – 0.77	-6.69	<0.001
Claude 3.5 Sonnet	0.70	0.63 – 0.78	-6.23	<0.001
Gemini 1.5 Pro	0.62	0.56 – 0.70	-8.34	<0.001

Gemini 1.5 Flash	0.64	0.57 – 0.71	-7.96	<0.001
Llama 3.2 3B	1.24	1.11 – 1.38	3.76	<0.001
Llama 405B	0.62	0.55 – 0.69	-8.51	<0.001

Random hatások

τ_{00} Mondat: item	0.17
τ_{00} item	0.09
τ_{00} rater_id	0.08
N mondat	2251
N item	7
N értékelő azonosító	441

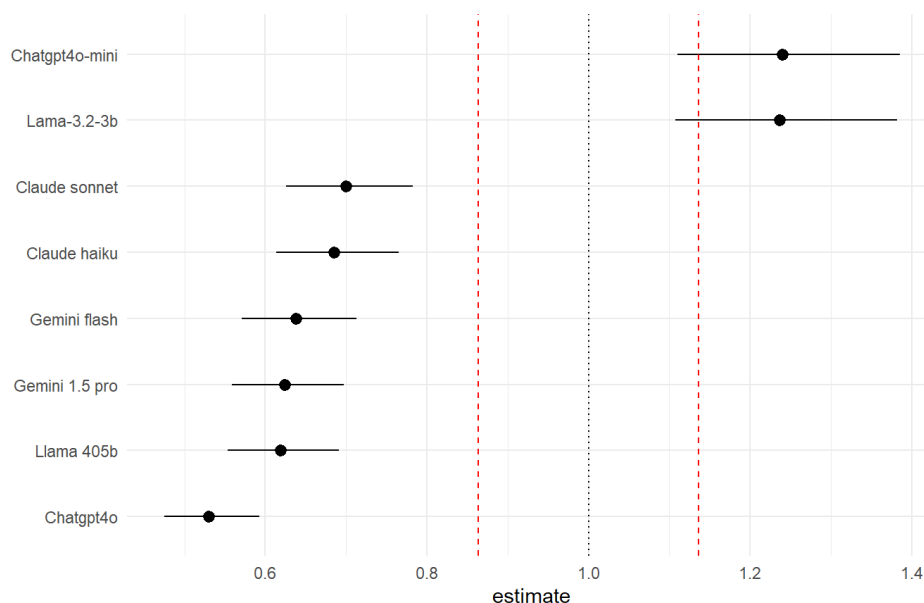
Megfigyelések száma	606131
---------------------	--------

AIC	1546976.209
-----	-------------

Megjegyzés: Az item az általunk adatgyűjtésre használt teszt itemeire utal.

A 0,5 pontnyi különbségként meghatározott legkisebb érdekes hatás multiplikatív skálára vetített értékei a 7. ábrán láthatók. A preregisztrált következtetési kritériumok alapján mindegyik modell szignifikánsan különbözött az emberi értékelések átlagától. Ez a 3. táblázatban közölt konfidencia intervallumokból is látható: egyik modell esetében sem voltak a konfidencia intervallumok a hipotézis igazolásához szükséges [0,864, 1,136] tartományban. A második hipotézisünket egyik modell esetén sem sikerült igazolnunk.

7. ábra: Modellek értékeléseinek átlaga az emberi értékelésekhez képest



Megjegyzés: A fekete pontok a nagy nyelvi modellek emberi értékelések átlagához képesti relatív arányát mutatják (rate ratio). Tehát, pl. a Claude sonnet model értékeléseinek átlaga az emberi értékelések átlagának a 70%-a, míg a piros sávok a preregisztrációban meghatározott SESOI intervallumot jelölik.

A harmadik hipotézisünket, mely szerint az emberi és a nagy nyelvi modellek értékeléseinek szórása között nincs szignifikáns különbség a második hipotézishez is alkalmazott location-scale modellel vizsgáltuk. A modell szóródásra vonatkozó jellemzői az alábbi táblázatban láthatók.

4. táblázat: A location-scale modell szóródásra vonatkozó jellemzői

Prediktorok	Regressziós együtthatók	CI	Statisztika	p-érték
(Intercept)	3.50	3.47 – 3.54	243.64	<0.001
ChatGPT-4o	1.69	1.66 – 1.71	69.88	<0.001
ChatGPT-4o mini	2.45	2.42 – 2.49	116.56	<0.001
Claude 3.5 Haiku	2.40	2.36 – 2.44	112.50	<0.001
Claude 3.5 Sonnet	1.68	1.65 – 1.71	68.03	<0.001
Gemini 1.5 Pro	5.70	5.60 – 5.81	187.23	<0.001
Gemini 1.5 Flash	5.51	5.41 – 5.61	187.86	<0.001

Llama 3.2 3B	0.93	0.91 – 0.94	-10.32	<0.001
Llama 405B	1.74	1.72 – 1.77	73.38	<0.001

Random hatások

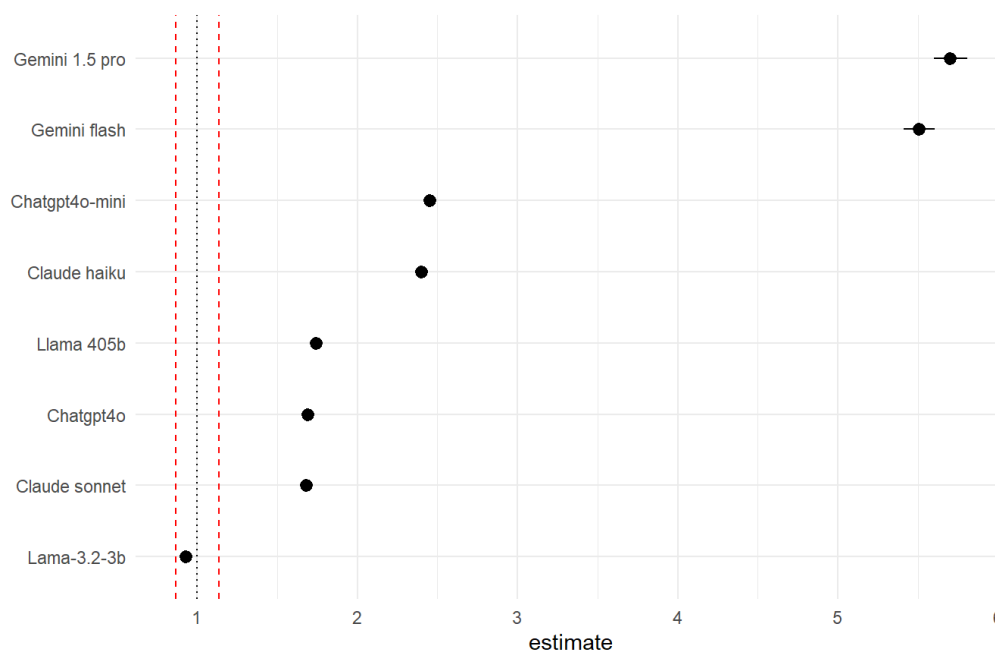
τ_{00} mondat:item	0.17
τ_{00} item	0.09
τ_{00} értékelő_id	0.08
N mondat	2251
N item	7
N értékelő_id	441

Megfigyelések száma	606131
---------------------	--------

AIC	1546976.209
-----	-------------

Az általunk meghatározott 0,5 pontnyi abszolút különbség multiplikatív skálán vett értékei a 8. ábrán piros vonallal láthatók. Az általunk vizsgált nagy nyelvi modellek közül egyedül a Llama 3.2 3B modell értéke került a legkisebb érdekes hatás (SESOI) preregisztrációban meghatározott tartományába (CI=0,91-0,94), így a hipotézisünket csak ennek a modellnek az esetén sikerült igazolnunk.

8. ábra: Modellek szóródásának (dispersion) vizsgálata az emberi értékelésekhez képest



Megjegyzés: A fekete pontok az AI modellek emberekhez képesti relatív arányát mutatják (rate ratio), míg a piros sávok a preregisztrációban meghatározott SESOI intervallumot jelölik.

Diszkusszió

Eredmények összefoglalása

Vizsgált hipotézisek

Dolgozatomban azt vizsgáltam, hogy a kutatás kezdetének időpontjában elérhető legfejlettebb és széles körben elterjedt nagy nyelvi modellek mennyire képesek az emberekhez hasonlóan értékelni a verbális agressziót. A verbális agresszió az online terekben különösen elterjedt (lásd például: Appel és mtsai., 2014), és ezáltal detektálása kulcsfontosságú lehet a további káros következmények elkerülésében, moderálásában. A számos kifejlesztett gépi tanuláson alapuló modell ellenére (del Bosque & Garza, 2014) jelenleg még nem áll rendelkezésünkre olyan modell, melyet széles körben használnának erre a célra. Kutatásom célkitűzése ennek megfelelően annak vizsgálata volt, hogy az elterjedt nagy nyelvi modellek mennyiben adnak az emberi értékelőkhöz hasonló eredményeket, és ezáltal mennyire alkalmazhatók a verbális agresszió detektálására.

A kutatás során vizsgált nagy nyelvi modellek közül a Llama 3.2 3B kivételével mindegyik 0,6 fölötti, erős korrelációt mutatott az emberi verbális agresszió értékelésekkel, míg a modellek egy részében ez az érték nagyon erős, 0,8 feletti volt (ChatGPT-4o mini, Claude 3.5 Sonnet, Gemini 1.5 pro). Ez a korrelációs eredmény azt mutatja, hogy a vizsgált nagy nyelvi modellek, és különösen a három kiemelt modell önmagában is konzisztens értékelést ad, és hatékonyan differenciál a mondatok között verbális agresszió szempontjából. Megfigyelhető továbbá, hogy a vizsgált modellek nagyobb verziói a ChatGPT-4o kivételével jobb korrelációs eredményt adtak, mint a kisebb verziók. A modellek komplexitásának növelése tehát a verbális agresszió értékelésében is jobb eredményekhez vezetett a legtöbb esetben.

A másik két hipotézisünket azonban, mely a nagy nyelvi modellek értékeléseinek átlagának és szórásának különbségeire vonatkozott az emberi értékelésekhez képest csak korlátozottan sikerült igazolnunk. A modellek mondatonkénti értékeléseinek átlaga egy modell esetén se volt az általunk meghatározott 0,5 pontnyi különbségen belül az emberi értékelésekhez képest. A szórás vizsgálata esetén pedig csak a Llama 3.2 3B modell mutatott

az emberi értékelésekhez hasonló szórást, (amely azonban a korrelációs vizsgálatban a legrosszabb eredményt érte el).

Az eredmények azt mutatják, hogy bár a modellek értékelései és az emberi értékelések között erős korreláció van, a modellek azonban az emberi értékelésektől eltérő átlagot és szórást eredményeztek. A szórás esetén a temperature értékek, (melyet a kutatásban az alapszinten hagytunk) változtatásával a válaszok variabilitása is befolyásolható, így elképzelhető, hogy egy másik temperature értékkel jobban megközelíthetőek lenne az emberi válaszok szórása.

Az eltérő átlag és szórás ellenére azonban a korrelációs eredmények alapján a nagy nyelvi modellek többsége és az emberi értékelők hasonlóan ítélték meg a mondatok egymáshoz viszonyított agresszivitását, (bár a nagy nyelvi modellek átlagosan alacsonyabb értékeléseket adtak, mint az emberek), így alkalmasak lehetnek a potenciálisan agresszív mondatok kiszűrésére.

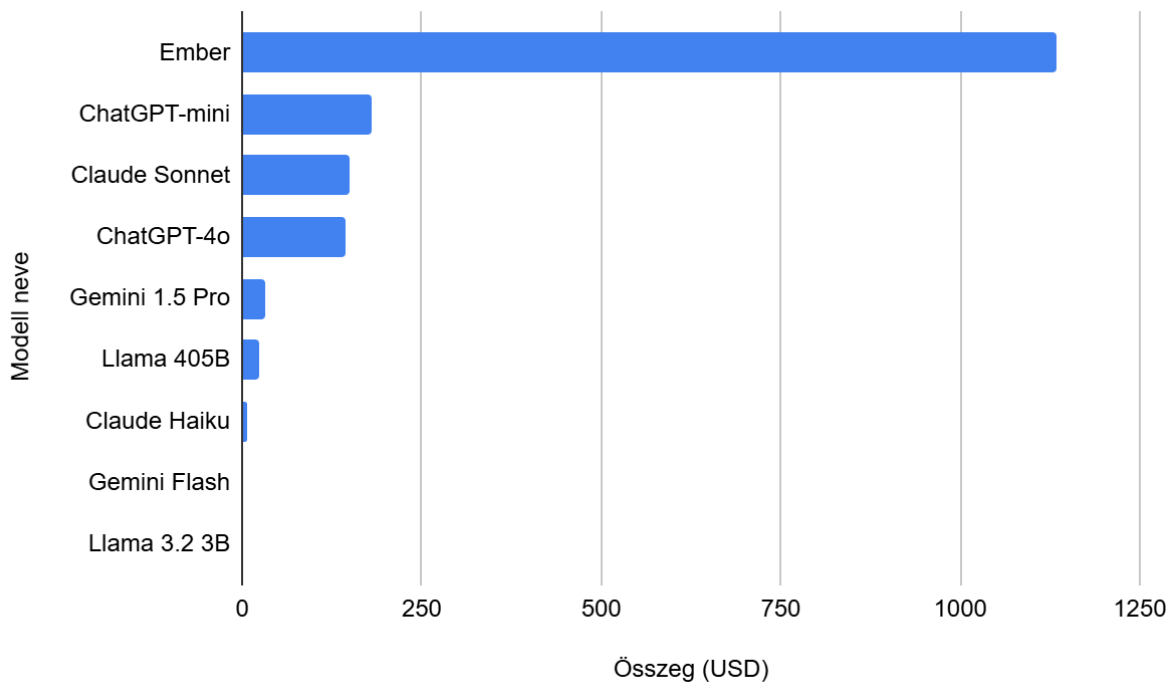
Hiányzó értékelések

A nagy nyelvi modellek értékeléseinek összegyűjtése során több modell is skálán nem értelmezhető értékeléseket adott, melyeket ki kellett zárunk. A ChatGPT-4o mini modell esetén ez adatgyűjtési hibából fakadt, (a feltöltött költségkeret az összes értékelés összegyűjtésének befejezése előtt lejárt, és a futtatás közbeni feltöltés problémákat okozhatott). A Llama 404B és a Claude 3.5 Sonnet esetén a válaszadás megtagadása azonban a modellbe épített korlátozó mechanizmusokra utal (Klyman, 2024), amely egyes modellek esetén kisebb válaszadási hajlandóságot eredményez szenzitív témákban, például verbális agresszió esetén (Brahman és mtsai., 2024). Kutatásunk alapján a Claude 3.5 Sonnet és a Llama 405B a kisebb válaszadási hajlandóság miatt csak korlátozottan alkalmazható a vizsgáltaknál agresszívebb mondatok detektálására.

Költségek

A kutatás egy másik fontos aspektusa a költségek vizsgálata. Az emberi értékelések összegyűjtése az időigényesség mellett költséges is. A kutatócsoportunk a toborzásra és a résztvevők fizetésére 914 angol fontot fordított, míg a nagy nyelvi modellek értékeléseinek költségei ennél jelentősen alacsonyabbak voltak. Az arányokat a 9. ábra szemlélteti. A legmagasabb, 0,8 feletti korrelációt elért modellek közül a Gemini 1.5 Pro bizonyult a legköltséghatékonyabbnak. A nagy nyelvi modellek használata az emberi értékelésekhez viszonyítva azonban minden esetben egy költséghatékony megoldást jelentett.

9. ábra: A kutatás során felmerült költségek modellek bontásában



Erősségek és limitációk

A kutatás az elsők között vizsgálta az ismert nagy nyelvi modellek teljesítményét egy olyan kérdésben, melynek eredménye a pszichológiai kutatások elősegítése mellett az online interakciók értékelésében is fontos szerepet játszik.

A kutatásnak egyik limitációja az összegyűjtött, értékelésre szánt mondatok többsége nem volt jelentősen agresszív. Elképzelhető, hogy a nagy nyelvi modellek értékelése máshogy alakult volna, ha az adathalmaz több, jelentősen agresszív mondatot is tartalmaz.

Ehhez szorosan kapcsolódik az is, hogy a nagy nyelvi modellek általában alacsonyabb értékelést adtak, mint az emberek. Ez két okra vezethető vissza. A modellek tanító adathalmazában valószínűleg az általunk vizsgáltaknál sokkal agresszívebb mondatok is szerepeltek, így a modellek feltehetően egy szélesebb nyelvi anyaghoz viszonyítva ítélték meg a valóban alacsony agresszivitás szintű mondatokat, míg az emberi értékelők sokkal inkább kontextuálisan, egymáshoz viszonyítva értékelték őket. Továbbá, az emberi értékelők esetén szerepet kaphatott egy kutatásnak való megfelelés hatása is (demand characteristics), amely növelhette az általuk adott pontokat, (a kérdés a mondatok agresszivitására kérdezett rá egyáltalán nem agresszív mondatok esetén is). Ez a két tényező is hozzájárulhatott ahhoz, hogy az emberi és a nagy nyelvi modellek értékelésének átlaga között szignifikáns különbség volt.

Ki kell emelni továbbá, hogy a verbális agresszió erősen nyelv és kultúrafüggő jelenség. Az emberi értékelők többsége angol anyanyelvű volt, ám elképzelhető, hogy más kultúrákban eltérő értékelések születtek volna. A kultúra általi meghatározottság is befolyásolhatta ezáltal az emberi és a rendkívül széles nyelvi anyagon tanított nagy nyelvi modellek értékelése közti különbséget.

Konklúzió

Összefoglalva, az általunk vizsgált nagy nyelvi modellek verbális agresszió értékelése erős korrelációt mutat az emberi értékelésekkel. Az átlagok különbségéből adódóan azonban a modellek jelenlegi formájukban nem alkalmazhatók arra, hogy az emberi értékelésekkel pontosan megegyező válaszokat adjanak. Ezt a továbbiakban egy eltérő instrukció alkalmazásával vagy a modellek továbbtanításával javítható. A magas korreláció miatt azonban erre nem feltétlenül van szükség, hiszen a modellek a saját értékelési rendszerükön belül így is az emberekhez hasonlóan, konzisztensen ítélték meg, a mondatok agresszivitását, így a verbális agresszió detektálására és értékelésére eredményesen alkalmazhatók.

Záró megjegyzések

Dolgozat befejezésének dátuma: 2025.01.08.

Nyilatkozat mesterséges intelligencia használatáról:

- Kijelentem, hogy a dolgozat szövegének megírásához mesterséges intelligencia segítségét nem vettem igénybe.
- A kutatás során az API hívást létrehozó Python kódhoz a Claude 3.5 Haiku és a Claude 3.5 Sonnet modelleket használtam kódgenerálásra és javításra (Anthropic, 2024, <https://claude.ai/new>). A pontos promptokat és válaszokat a modell memóriájából fakadó korlátok miatt nem tudom beilleszteni, mert azok a dolgozat megírásakor már törölték.

Irodalomjegyzék

Anthropic. (2025). *Claude*. (Jan 8 version) [Large language model]. <https://claude.ai/new>

Appel, M., Stiglbauer, B., Batinic, B., & Holtz, P. (2014). Internet use and verbal aggression: The moderating role of parents and peers. *Computers in Human Behavior*, 33,

- 235–241. <https://doi.org/10.1016/j.chb.2014.01.007>
- Basile, P., de Gemmis, M., Musacchio, E., Polignano, M., Semeraro, G., Siciliani, L., Tamburrano, V., Barletta, V., Caivano, D., Battista, F., Curci, A., Scardigno, R., Calvano, G., & Sorianello, P. (2024). *Explaining Intimate Partner Violence with LLaMAntino*.
- Bauer, N., Preisig, M., & Volk, M. (2024). Offensiveness, Hate, Emotion and GPT: Benchmarking GPT3.5 and GPT4 as Classifiers on Twitter-specific Datasets. In R. Kumar, A. Kr. Ojha, S. Malmasi, B. R. Chakravarthi, B. Lahiri, S. Singh, & S. Ratan (Szerk.), *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024* (o. 126–133). ELRA and ICCL. <https://aclanthology.org/2024.trac-1.14>
- Brahman, F., Kumar, S., Balachandran, V., Dasigi, P., Pyatkin, V., Ravichander, A., Wiegrefe, S., Dziri, N., Chandu, K., Hessel, J., Tsvetkov, Y., Smith, N. A., Choi, Y., & Hajishirzi, H. (2024). *The Art of Saying No: Contextual Noncompliance in Language Models* (arXiv:2407.12043). arXiv. <https://doi.org/10.48550/arXiv.2407.12043>
- Brooks, M., Bolker, B., Kristensen, K., Maechler, M., Magnusson, A., McGillicuddy, M., Skaug, H., Nielsen, A., Berg, C., Benth, K. van, Sadat, N., Lüdtke, D., Lenth, R., O'Brien, J., Geyer, C. J., Jagan, M., Wiernik, B., Stouffer, D. B., & Agronah, M. (2024). *glmmTMB: Generalized Linear Mixed Models using Template Model Builder* (Verzió 1.1.10) [Software]. <https://cran.r-project.org/web/packages/glmmTMB/index.html>
- Cole, D. A., Martin, N. C., Sterba, S. K., Sinclair-McBride, K., Roeder, K. M., Zelkowitz, R., & Bilsky, S. A. (2014). Peer victimization (and harsh parenting) as developmental correlates of cognitive reactivity, a diathesis for depression. *Journal of Abnormal Psychology*, 123(2), 336–349. <https://doi.org/10.1037/a0036489>
- DeepL Translate: *The world's most accurate translator*. (é. n.). Elérés 2025. január 6., forrás <https://www.deepl.com/en/translator>
- Del Bosque, L. P., & Garza, S. E. (2014). Aggressive Text Detection for Cyberbullying. In A. Gelbukh, F. C. Espinoza, & S. N. Galicia-Haro (Szerk.), *Human-Inspired Computing and Its Applications* (o. 221–232). Springer International Publishing. https://doi.org/10.1007/978-3-319-13647-9_21
- Fireworks—*Fastest Inference for Generative AI*. (é. n.). Elérés 2025. január 6., forrás <https://fireworks.ai/>

- Google (2025). *Gemini*. (Jan 8 version) [Large language model]. <https://gemini.google.com/app>
- Hanley, A. J., & Gibb, B. E. (2011). Verbal Victimization and Changes in Hopelessness Among Elementary School Children. *Journal of Clinical Child & Adolescent Psychology*, 40(5), 772–776. <https://doi.org/10.1080/15374416.2011.597086>
- Hengel, P., & Andringa, T. C. (2007). *Verbal aggression detection in complex social environments*. 15–20. <https://doi.org/10.1109/AVSS.2007.4425279>
- Hester, J., Wickham, H., Csárdi, G., library), libuv project contributors (libuv, Joyent, library), I. and other N. contributors (libuv, Software, P., & PBC. (2024). *fs: Cross-Platform File System Operations Based on „libuv”* (Verzió 1.6.5) [Software]. <https://cran.r-project.org/web/packages/fs/index.html>
- Klyman, K. (2024). Acceptable Use Policies for Foundation Models. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7, 752–767. <https://doi.org/10.1609/aies.v7i1.31677>
- Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., Kocoń, A., Koptyra, B., Mieszczenko-Kowszewicz, W., Miłkowski, P., Oleksy, M., Piasecki, M., Radliński, Ł., Wojtasik, K., Woźniak, S., & Kazienko, P. (2023). ChatGPT: Jack of all trades, master of none. *Information Fusion*, 99, 101861. <https://doi.org/10.1016/j.inffus.2023.101861>
- Meta (2025). (Jan 8 version) [Large language model]. <https://www.llama.com/>
- Lüdecke, D., Bartel, A., Schwemmer, C., Powell, C., Djalovski, A., & Titz, J. (2024). *sjPlot: Data Visualization for Statistics in Social Science* (Verzió 2.8.17) [Software]. <https://cran.r-project.org/web/packages/sjPlot/index.html>
- Lüdecke, D., Makowski, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., Wiernik, B. M., Thériault, R., Arel-Bundock, V., Jullum, M., gjo11, Bacher, E., & Luchman, J. (2024). *performance: Assessment of Regression Models Performance* (Verzió 0.12.4) [Software]. <https://cran.r-project.org/web/packages/performance/index.html>
- Makowski, D., Wiernik, B. M., Patil, I., Lüdecke, D., Ben-Shachar, M. S., Thériault, R., White, M., & Rabe, M. M. (2024). *correlation: Methods for Correlation Analysis* (Verzió 0.8.6) [Software]. <https://cran.r-project.org/web/packages/correlation/index.html>
- Nina-Gutiérrez, E. A., Pacheco-Alanya, J. E., & Morales-Arevalo, J. C. (2024). Multilingual Detection of Cyberbullying on Social Networks Using a Fine-Tuned GPT-3.5 Model.

- In A. Fred, A. Hadjali, O. Gusikhin, & C. Sansone (Szerk.), *Deep Learning Theory and Applications* (o. 252–263). Springer Nature Switzerland.
https://doi.org/10.1007/978-3-031-66705-3_17
- OpenAI. (2025). *ChatGPT* (Jan 8 version) [Large language model].
<https://chat.openai.com/chat>
- Open Science Framework* (é.n.) Elérés: 2025.01.08., forrás <https://osf.io/tpzw9>
- Poling, D. V., Smith, S. W., Taylor, G. G., & Worth, M. R. (2019). Direct verbal aggression in school settings: A review of the literature. *Aggression and Violent Behavior*, 46, 127–139. <https://doi.org/10.1016/j.avb.2019.01.010>
- Prolific* | *Quickly find research participants you can trust.* (é. n.). Elérés 2024. július 4., forrás <https://www.prolific.com/>
- R Core Team (2024). *_R: A Language and Environment for Statistical Computing_*. R Foundation for Statistical Computing, Vienna, Austria.
<https://www.R-project.org/>
- Robinson, D., Hayes, A., Couch [aut, S., cre, Software, P., PBC, Patil, I., Chiu, D., Gomez, M., Demeshev, B., Menne, D., Nutter, B., Johnston, L., Bolker, B., Briatte, F., Arnold, J., Gabry, J., Selzer, L., Simpson, G., ... Reinhart, A. (2024). *broom: Convert Statistical Objects into Tidy Tibbles* (Verzió 1.0.7) [Software].
<https://cran.r-project.org/web/packages/broom/index.html>
- Rosenzweig, S., Ludwig, D. J., & Adelman, S. (1975). Retest Reliability of the Rosenzweig Picture-Frustration Study and Similar Semiprojective Techniques. *Journal of Personality Assessment*, 39(1), 3–12. https://doi.org/10.1207/s15327752jpa3901_1
- Rueger, S. Y., & Jenkins, L. N. (2014). Effects of peer victimization on psychological and academic adjustment in early adolescence. *School Psychology Quarterly*, 29(1), 77–88. <https://doi.org/10.1037/spq0000036>
- Sadiq, S., Mehmood, A., Ullah, S., Ahmad, M., Choi, G. S., & On, B.-W. (2021). Aggression detection through deep neural model on Twitter. *Future Generation Computer Systems*, 114, 120–129. <https://doi.org/10.1016/j.future.2020.07.050>
- Trillo, J. R., Cabrerizo, F. J., Pérez, I. J., Morente-Molinera, J. A., & Herrera-Viedma, E. (2024). A new Consensus Reaching Method for Group Decision-Making based on the Large Language Model Gemini for Detecting Hostility During the Discussion Process. *2024 IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, 1–8. <https://doi.org/10.1109/EAIS58494.2024.10570029>
- Waring, E., Quinn, M., McNamara, A., Rubia, E. A. de la, Zhu, H., Lowndes, J., Ellis, S.,

- McLeod, H., Wickham, H., Müller, K., RStudio, functions), I. (Spark, Kirkpatrick, C., Brenstuhl, S., Schratz, P., Ibusett, Korpela, M., Thompson, J., McGehee, H., ... Stewart, H. M. (2022). *skimr: Compact and Flexible Summaries of Data* (Verzió 2.1.5) [Software]. <https://cran.r-project.org/web/packages/skimr/index.html>
- Wen, Z., & Younes, R. (2024). *ChatGPT v.s. Media Bias: A Comparative Study of GPT-3.5 and Fine-tuned Language Models* (arXiv:2403.20158). arXiv. <https://doi.org/10.48550/arXiv.2403.20158>
- Wickham, H., & RStudio. (2023). *tidyverse: Easily Install and Load the „Tidyverse”* (Verzió 2.0.0) [Software]. <https://cran.r-project.org/web/packages/tidyverse/index.html>