Eötvös Loránd Tudományegyetem

Társadalomtudományi Kar

MESTERKÉPZÉS

A nagy nyelvi modellek alkalmazási lehetőségei a szöveganalitikai annotációban

Konzulens:

Katona Eszter Rita

Készítette:

Fodor Alexandra

HMGXSF

Survey Statisztika és Adatanalitika szak

2025. április

EREDETISÉGNYILATKOZAT

(Kitöltés után a szakdolgozat/diplomamunka részét képezi.)

Alulírott…… Fodor Alexandra ..... *[a hallgató neve]* az ELTE TáTK ……..Survey Statisztika és Adatanalitika MSc..... *[a hallgató szakja]* hallgatója fegyelmi felelősségem tudatában nyilatkozom, és aláírásommal igazolom, hogy a(z) ..... A nagy nyelvi modellek alkalmazási lehetőségei a szöveganalitikai annotációban .... *[a szakdolgozat/diplomamunka címe]* című szakdolgozat/diplomamunka **saját, önálló szellemi munkám,** az abban hivatkozott, nyomtatott és elektronikus szakirodalom felhasználása a szerzői jogok szabályainak megfelelően történt.

Tudomásul veszem, hogy szakdolgozat/diplomamunka esetén plágiumnak számít:

- a szó szerinti idézet vagy fordítás közlése idézőjel és hivatkozás megjelölése nélkül;
- a tartalmi idézet hivatkozás megjelölése nélkül;
- más publikált gondolatainak saját gondolatként való feltüntetése.

Alulírott kijelentem, hogy a plágium fogalmát, valamint a HKR 74/A–C. §-aiban foglalt rendelkezéseket megismertem, és tudomásul veszem, hogy

- a szakdolgozatom/diplomamunkám szövege a benyújtása után plágiumellenőrző szoftverrelvizsgálható,
- plágium esetén szakdolgozatom/diplomamunkám értékelés nélkül visszautasításra kerül,
- plágiumvétség esetén fegyelmi eljárás indítható.

Budapest/Szombathely, 2025. április 13. .....

.....Fodor Alexandra …..

.....

a hallgató aláírás

Absztrakt

A szakdolgozatomban a generatív nagy nyelvi modellek szöveganalitikai annotációs feladatokban való alkalmazhatóságát vizsgáltam egy depresszióval kapcsolatos szövegkorpusz segítségével. A kutatásom során a zárt forráskódú GPT-4o mini és a nyílt forráskódú Llama 3.3 70B teljesítményét hasonlítottam össze. Mindkét modell esetében kitértem a zero-shot és a few-shot technikák eredményeinek összevetésére is. A pontosság tekintetében a few-shot megközelítés enyhe javuláshoz vezetett a zero-shot technikával szemben. A Llama modell pedig összeségében kicsivel jobb teljesítményt nyújtott, mint a GPT. A két modell pontosság tekintetében inkább közepes eredményt nyújtott, ezzel szemben a konzisztenciájuk és megbízhatóságuk magasnak mondható.

Kulcsszavak: szövegannotáció, nagy nyelvi modell, LLM, GPT, Llama, NLP, szöveganalitika, klasszifikáció, prompt, zero-shot, few-shot, depresszió

Legfontosabb irodalmak:

Németh, R., Máté, F., Katona, E., Rakovics, M., & Sik, D. (2022). Bio, psycho, or social: Supervised machine learning to classify discursive framing of depression in online health communities. *Quality & Quantity, 56*, 3933–3955. DOI: https://doi.org/10.1007/s11135-021-01299-0

Németh, R., Sik, D., & Máté, F. (2020). Machine learning of concepts hard even for humans: The case of online depression forums. *International Journal of Qualitative Methods, 19*, 1–8. DOI: https://doi.org/10.1177/1609406920949338

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences, 120*(30), e2305016120. DOI: https://doi.org/10.1073/pnas.2305016120

Tartalomjegyzék

1. Bevezetés

A természetes-nyelvfeldolgozás különböző módszerei nagy mennyiségű szöveges adat elemzését teszik lehetővé. Ezek közé az adatforrások közé tartoznak többek között a különböző cikkek, posztok és kommentek is. A szöveges adatok elemzése, azonban extenzív adattisztítással és adatfeldolgozással jár. A nyelvi sajátosságok és a kontextus függő jelentések miatt az automatikus feldolgozás különösen sok kihívást rejt magában. Ebből adódóan számos olyan módszer és kutatási terület van, amely esetében egy felcímkézett adatkorpusz létrehozására is szükség van. Az ilyen annotált adathalmazok elengedhetetlenek a felügyelt gépi tanulási modellek tanításához. Ide tartozik például az emócióelemzés, amely esetében a szövegekben megjelenő érzelmek azonosítása a célunk (Demszky et al., 2020), a névelem-felismerés, amely célja a szövegben szereplő tulajdonnevek azonosítása (Bontcheva et al., 2017), és szintén ilyen a szövegek különböző csoportokba történő klasszifikációja is (Qureshi & Sabih, 2021). Az egyes eljárások implementációjához azonban nagy mennyiségű és pontos címkékkel ellátott adathalmazra van szükség.

Az annotált korpusz létrehozásában rendszerint emberi kódolók vesznek részt, a folyamat azonban nagy mennyiségű időt vesz igénybe, emellett pedig magas költségekkel is jár. Különösen azt figyelembe véve, hogy a pontosság növelése érdekében egy szöveget általában minimum két ember annotál és az ellenkező vélemények esetében arra is szükség lehet, hogy egy fő annotátor döntsön a kérdésben (Németh, Sik & Máté, 2020). Szintén nem elhanyagolható kérdéskör az annotátorok megfelelő betanítása, illetve hogy mennyire képesek elfogulatlan döntéseket hozni. Emellett érdemes megemlíteni, hogy számos olyan témakör van, amelyek feldolgozása érzelmileg megterhelő lehet az emberi annotátorok számára, például gyűlöletbeszédet tartalmazó posztok és kommentek (De Gibert et al., 2018).

A fenti okok és az utóbbi évek technológiai fejlődésének következtében egyre több figyelem irányul arra, hogy milyen felhasználási lehetőségei vannak a különböző nagy nyelvi modelleknek (Large Language Model – LLM) a szöveganalitikában és azon belül is az annotációban (Gilardi et al., 2023; Goel et al., 2023; Labruna et al., 2023; Ostyakova et al., 2023). Az ilyen modellek képesek lehetnek az emberi annotáció kiváltására vagy legalább annak támogatására, gyorsítva a folyamatokat és csökkentve az erőforrásigényt. A szakdolgozatom célja egy gyakorlati példán keresztül bemutatni a LLM-ek annotációban

történő alkalmazásának módszerét. Ehhez egy, a karon már jól ismert depresszióval kapcsolatos adatbázist használok majd, amely különböző online hozzászólásokat és fórumbejegyzéseket tartalmaz. A klasszifikáció során a cél annak eldöntése, hogy az adott bejegyzés írója biomedikális, pszichológiai vagy szociológia szempontból keretezi a depressziót.

A fő kutatási kérdésem, hogy a nagy nyelvi modellek mennyire képesek pontosan reprodukálni az emberi annotátorok által adott címkéket és hogy mindezt mekkora konzisztencia mellett képesek végrehajtani. Ehhez a már fentiekben említett adatbázis nagyjából 4500 felcímkézett kommentjét és posztját használom majd fel. A dolgozatomban kitérek arra is, hogy ezek a modellek mennyire képesek kezelni azt, ha a kategóriák nem egymást kizáróak és egy szöveghez több címke is tartozhat (azaz mennyire alkalmasak multi-label klasszifikációra). Emellett a vizsgálat során figyelmet fordítok arra is, hogy a modellek hogyan teljesítenek az egyértelmű és a nehezebben besorolható eseteknél. Nehezen besorolhatónak azokat a szövegeket nevezzük, ahol a két annotátor elsődleges címkéi eltérőek voltak. Célom ezen felül összehasonlítani a különböző modellek és promptolási eljárások eredményeit.

A dolgozatomban először részletesen bemutatom a releváns szakirodalmat. Ezen belül kitérek a szöveganalitika általános elméleti hátterére és az annotáció szerepére a jellemző módszereken belül. Emellett részletesen ismertetem a különböző nagy nyelvi modellek jellegzetességeit is. Végül pedig kitérek arra is, hogy az eddigi kutatások miként és milyen eredmények mellett használták fel a nagy nyelvi modelleket a szöveganalitikai annotációban.

A harmadik fejezetben részletesen bemutatom a felhasznált adatbázis jellemzőit, létrejöttének módját, illetve a korpusszal kapcsolatos korábbi kutatások eredményeit.

A szakdolgozatom negyedik fejezetében ismertetem a kutatásom módszertanát. Bemutatom a vizsgálatomban szereplő nagy nyelvi modelleket, az annotációhoz felhasznált promptokat, illetve azokat a mérőszámokat, amelyek segítségével kiértékelem a kutatásban

részt vevő modellek teljesítményét. Ezen felül kitérek a temperature[1] paraméter hangolására és ennek hatására is.

Ezt követően részletesen bemutatom a kutatásom eredményeit. Összehasonlítom a különböző modellek eredményeit, ennek során egyaránt kitérek ezek pontosságára és konzisztenciájára. A pontosságot globálisan és kategóriák szerint is megvizsgálom. A konzisztenciát és a megbízhatóságot pedig azonos és eltérő modellek esetében is.

Végül kitérek a különböző korlátokra és limitációkra, majd összegzem a dolgozatom eredményeit és felvázolom a témához kapcsolódó további kutatási lehetőségeket.

2. Szakirodalmi áttekintés

2.1. Szöveganalitika és annotáció

A természetes nyelvfeldolgozás (NLP) célja, hogy számítógépes módszerekkel elemezze és értelmezze az emberi nyelvet. A kezdeti kutatások főként a nyelvi szerkezetek automatikus feldolgozására és olyan alapvető technológiák kifejlesztésére irányultak, mint a gépi fordítás, a beszédfelismerés és a beszédszintézis. A kutatók mostanra továbbfejlesztették ezeket az eszközöket, és széles körben alkalmazzák őket gyakorlati problémák megoldására is (Hirschberg & Manning, 2015).

A természetes nyelvfeldolgozásban két nagy szemléletmódot különböztethetünk meg, ezek a felügyelt és a felügyelet nélküli tanítási módszerek. A felügyelet nélküli technikák esetében rendszerint nem rendelkezünk előzetes információval a szövegek kapcsán, nincsnek felcímkézett adataink. Ebben az esetben a cél általában valamilyen látens tartalom feltárása (Németh, Katona & Kmetty, 2020). Az egyik legelterjedtebb módszer a topikmodellezés, például a LDA (Latent Dirichlet Allocation) algoritmus, amely különböző témákba rendezi a dokumentumokat (Blei et al., 2003). Ezen kívül szintén elterjedtek a különböző klaszterezési eljárások (Aggarwal & Zhai, 2012b) és szóbeágyazási modellek is (Mikolov et al., 2013).

Ezzel szemben a felügyelt tanítási módszerek esetében vannak előzetes elképzeléseink és információink a vizsgált szövegekről, például rendelkezünk egy már felcímkézett

[1] A temperature paraméter a generálás során a tokenek kiválasztásának valószínűségi eloszlását módosítja: alacsony értéknél a modell inkább a legvalószínűbb tokeneket választja, míg magasabb értéknél nagyobb eséllyel választ ritkább, így kreatívabb szavakat is (Jurafsky & Martin, 2025).

szövegkorpusszal a célunk pedig az, hogy a címkézetlen eseteket előre meghatározott kategóriákba soroljuk be minél pontosabban. A legmegfelelőbb modell illesztéséhez és kiválasztásához először két részre osztjuk a felcímkézett adatbázisunkat, egy tanító- és egy teszthalmazra. A tanítóhalmazon illesztjük a modellünket, a teszthalmazon pedig kiértékeljük annak pontosságát (Németh, Katona & Kmetty, 2020). Ilyen felügyelt tanítási módszerek például a döntési fák, a regressziós és naiv bayes modellek, a neurális hálók és a support vector machine (SVM) is (Aggarwal & Zhai, 2012a).

A felügyelt tanításban tehát egy kiemelten fontos lépés a felcímkézett szövegkorpusz létrehozása. Ez a gyakorlatban a legtöbbször emberi annotátorok által történik meg. Mivel az egyes modellek a felcímkézett adatokon tanulnak, ezért elengedhetetlen, hogy ezek a lehető legjobb minőségűek legyenek (Németh, 2024). Hovy és Lavid (2010) meglátásai szerint ennek érdekében az annotációs folyamatnak világos utasításokon kell alapulnia. Első lépésként pontosan meg kell határozni, milyen jelenségeket kívánunk vizsgálni, és milyen kategóriák szerint osztályozzuk az adatokat. Emellett fontos az annotáció céljának tisztázása is. Az annotáció megbízhatóságát növeli, ha az annotátorok megfelelő képzésben részesülnek, és a címkézést egymástól függetlenül legalább két annotátor végzi el. Az esetleges eltéréseket ellenőrzött módon kell megbeszélni, hogy a döntések következetesek maradjanak. Végezetül pedig elengedhetetlen az annotációk minőségének kiértékelése, amelyet a kutatók leggyakrabban az annotátorok közötti egyetértés megvizsgálásával valósítanak meg. Erre a leggyakrabban használt mutatók a Cohen-féle Kappa (Cohen, 1960) és a Krippendorff-féle alpha (Krippendorff, 2004), előbbiről részletesebben is beszámolok majd a szakdolgozatom módszertani fejezetében.

Mivel a tanítóhalmaz mérete pozitívan befolyásolja a betanított modellek pontosságát, ezért a kutatók gyakran törekednek akár több tízezer adatpontból álló korpusz létrehozására. Az annotáció azonban idő- és költségigényes folyamat, így a szakértői annotáció helyett elterjedt az olyan crowdsourcing platformok használata, mint a Mechanical Turk[2] vagy a korábban CrowdFlower-ként ismert Figure Eight, amely ma már Appen néven fut[3]. Ezeken az oldalakon a kutatók tulajdonképpen különböző annotációs feladatokat adhatnak meg, amelyeket a platformon bedolgozó, gyakran nem szakértői annotátorok végeznek el (Németh,

[2] Amazon Mechanical Turk: https://www.mturk.com/
[3] Appen: https://www.appen.com/

2024). Snow és munkatársai (2008) kutatásukban arra a következtetésre jutottak, hogy az MTurk platformon keresztül toborzott nem szakértői annotátorok is képesek jó minőségű címkézést végezni, különösen akkor, ha több annotátor válaszait kombinálják. A több, nem szakértői annotátor címkéiből tanított modellek, pedig egyes esetekben jobban is teljesítettek, mint egy szakértő címkéiből tanított modellek, amely arra enged következtetni, hogy még a szakértők esetében is fellelhető bizonyos mértékű torzítás, amelyet több ember véleményének átlagolása képes csökkenti. Hasonló megállapításra jutott Aroyo és Welty (2015) is, akik szerint az annotáció egy mítoszának is nevezhető az a gondolat, miszerint szakértői annotátorok jobbak lennének, mint a nem szakértők.

Látható tehát hogy az annotációnál fontosabb, hogy az annotációt több független kódoló végezze, ezáltal több különböző gondolkodásmód jelenjen meg, mint a tárgyi tudás. Ezért is lehet sikeres a crowdsourcing alapú annotálás annak ellenére, hogy az annotációs folyamat során azonban számos formában jelentkezhet torzítás, amelyek befolyásolhatják az annotált adatok minőségét és ezzel együtt a gépi tanulási modellek teljesítményét is. Geva és munkatársai (2019) kutatásukban arra mutattak rá, hogy a crowdsourcing annotáció során torzításhoz vezethet az, ha az annotátorok egy kis része címkézi a szövegek jelentős részét. Ezen felül a társadalmi és kulturális torzítások szintén befolyásolhatják az annotációt. Sap és munkatársai (2019) kutatásukban például kimutatták, hogy az annotáció során az annotátorok személyes és kulturális háttere hatással lehet az érzelmi töltet megítélésére vagy egyes kifejezések értelmezésére. Ennek következményeként bizonyos csoportok szövegei következetesen eltérő módon lehetnek címkézve, ami torzítja a modellek tanulási folyamatait és azok későbbi döntéseit is.

Összegzésül, az annotáció minősége kulcsfontosságú a gépi tanulási modellek sikeressége szempontjából, mivel a címkézett adatok meghatározzák a modellek tanulási alapját. Az annotáció során azonban számos különböző szempontot érdemes figyelembe venni. Ebből adódóan a folyamat költség- és időigényes lehet, valamint a legnagyobb körültekintés mellett is felmerülhetnek torzítások. A nagy nyelvi modellek alkalmazása érdekes lehetőséget kínál az annotációs folyamatok automatizálására, csökkentve az idő- és költségigényt, a torzítás kérdéskörét azonban ebben az esetben is érdemes vizsgálni, hiszen ezen modellek működési folyamata is nagy mértékben függ azoktól az adatoktól, amelyeken betanították őket, ez pedig szintén magában foglalhatja a torzítás lehetőségét úgy, ahogy az emberi annotáció is.

2.2. A nagy nyelvi modellek működése és jellegzetességei

A mesterséges intelligencia területén belül az utóbbi években egyre több nagy nyelvi modell jelent meg, amelyeket ezzel együtt egyre szélesebb körben kezdtek el használni is. Ezek a modellek tulajdonképpen nagyon nagy méretű neurális hálók, amelyek igen nagy szövegkorpuszon, mélytanulási technikákkal kerültek betanításra (LeCun et al., 2015).

A legegyszerűbb esetben a nagy nyelvi modellek bemenete szöveg, amit először számszerű formába kell hozni. Az adat-előfeldolgozás része a tokenizálás, azaz a szöveg feldarabolása kisebb egységekre, ezek a tokenek lehetnek szavak, szótagok vagy karaktertömbök is. A tokeneket ezután a modell beágyazott vektorként (embedding) reprezentálja, így a neurális hálózat numerikus formában kapja meg a szöveget.

A mai nagy nyelvi modellek jelentős része a transformer neurális hálózati architektúrán alapul. A transformer modellek legfőbb újítása a figyelemmechanizmus (attention mechanism), amely lehetővé teszi, hogy a modell egy adott token feldolgozásakor ne csak annak közvetlen szomszédos szavait, azaz lokális kontextusát vegye figyelembe, hanem a bemenet távolabbi részeire is fókuszálhasson. Ez a gyakorlatban azt jelenti, hogy a modell képes összefüggéseket felismerni akár olyan szövegrészek között is, amelyek fizikailag messze helyezkednek el egymástól a szövegben. Például a dokumentum elején szereplő mondat értelmezéséhez felhasználhatja a dokumentum végén található információkat is. Ez a képesség különösen hasznos hosszabb szövegek feldolgozásánál, mivel ennek köszönhetően a hálózat hatékonyan tanulja meg az úgynevezett „hosszú távú összefüggéseket” a szövegben. A transformer rétegekben önfigyelési (self-attention) mechanizmus működik, több párhuzamos figyelemfejjel (multi-head attention), amelyek eltérő szempont szerint mérik fel, hogy egy-egy token mennyire releváns a többi token kontextusában. A transformer megközelítés további előnye, hogy teljes mértékben párhuzamosítható, így nagyon nagy adathalmazokon is hatékonyan tanítható (Vaswani et al., 2017).

A generatív nagy nyelvi modellek működésének egy másik lényegi eleme az autoregresszív előrejelzés. A modellt úgy tanítják be, hogy mindig a már meglévő tokenek alapján próbálja megjósolni a következő tokent.

A nagy nyelvi modellek tanítása általában több lépcsőből áll. Először egy előképzési (pre-training) folyamat során nagyon nagy méretű szövegkorpuszon egy felügyelet nélküli tanítás

zajlik például autoregresszív célfügvénnyel. A modell ebben a fázisban még nem konkrét feladatot tanul, hanem általános nyelvi mintázatokat, szabályszerűségeket sajátít el. Ilyen előképzési adathalmaz lehet például az Elérhető fellelhető szövegek milliárdjai, Wikipédia cikkek, könyvek, fórumok tartalma stb. Például a GPT-3 modell megközelítőleg 175 milliárd paraméterrel rendelkezik, és több száz milliárd tokenből álló korpuszon lett előtanítva. Az ezt követő szakaszban, a finomhangolás (fine-tuning) során már egy felügyelt tanítás zajlik címkézett adatokkal, amely során specifikus feladatokra optimalizálják a modellt (Brown et al., 2020; Jurafsky & Martin, 2025).

A finomhangolás egy célja lehet az is, hogy a modell minél pontosabban kövesse a számára adott utasításokat. Az OpenAI kutatói ehhez egyrészt példamutató párbeszédeket, másrészt egy megerősítéses tanulást alkalmaztak emberi visszajelzéssel (reinforcement learning from human feedback, RLHF). Ennek keretében emberek rangsorolták a modell által adott kimeneteket, és ezek alapján egy jutalmazó függvényt tanítottak, majd ezzel finomhangolták a modellt úgy, hogy a magasabb értékelésű válaszok irányába tereljék a generálást. (Ouyang et al., 2022).

Mivel a nagy nyelvi modellek előképzés során nagy nyelvi tudást gyűjtenek össze, így egyetlen alapmodell (foundation model) számos különböző feladatra használható. A modellek viszonylag kevés finomhangolással vagy akár csak megfelelő utasításokkal (promptokkal) rávehetők egy adott feladat megoldására (Bommasani et al., 2021). A nagy nyelvi modellek így képesek lehetnek akár arra is, hogy zero-shot vagy few-shot megközelítéssel oldjanak meg különböző feladatokat külön finomhangolás nélkül. Zero-shot esetben a modell csak a feladat leírását és az elvégzéshez szükséges utasításokat kapja meg, míg few-shot esetben különböző példákat is elérhetővé tesznek a részükre, amelyek segítik a feladat értelmezését (Brown et al., 2020). Ilyen feladatok közé tartozik például különböző szövegek összegzése, lefordítása, generálása vagy klasszifikációja különböző kategóriákba, ezutóbbival foglalkozom én is a szakdolgozatomban.

Ugyanakkor a nagy nyelvi modelleknek is megvannak a hátrányaik, ezek közül az egyik a bennük rejlő torzítás lehetősége. Caliskan és munkatársai (2017) tanulmányukban kimutatták, hogy egy egyszerű word embedding modell is emberi jellegű előítéleteket hordozhat, például a női neveket a “család” és “otthon” koncepciókhoz közelebbinek találta, míg a férfineveket a “karrier” és “szakma” fogalmakhoz. Ezek az előítéletek abból fakadnak, hogy a modell a

társadalomban előforduló asszociációkat tanulta meg a szövegek alapján, vagyis a nyelv magában hordozza a társadalmi és kulturális elfogultságokat, amelyeket a gépi tanulás felerősíthet. Habár a nagy nyelvi modellek tanításához használt szövegek száma egyre növekszik, a modellek továbbra is átvehetik azokat az előítéleteket és torzításokat, amelyek a tanításukhoz használt szövegekben is fellelhetők. Bender és munkatársai (2021) arra figyelmeztetnek, hogy a hatalmas korpuszok megfoghatatlan tartalma miatt a modellek olyan váratlan vagy káros szövegeket is előállíthatnak, amelyek a fejlesztők számára átláthatatlan módon kerültek bele a rendszerbe. A szerzők sztochasztikus papagáj metaforával élnek, a modell a tanult statisztikai mintázatokat értelem nélkül utánozza, így akaratlanul is megismételheti a tréningadatokban lévő problémás tartalmakat és torzításokat.

A transformer neurális hálók és ezzel együtt a nagy nyelvi modellek egy másik, kevésbé szembetűnő hátránya, hogy tanításuk és használatuk során egyaránt jellemző rájuk, hogy számításigényes mivoltukból adódóan nagy az energiafogyasztásuk, ezáltal pedig az ökológiai lábnyomuk is. Különösen igaz ez, ha figyelembe vesszük, hogy az utóbbi hét évben a nagy nyelvi modellek fejlődésével a tanításukhoz használt korpuszok mérete és a modellek paramétereinek száma egyaránt jelentős növekedést mutatott. Ebből adódóan az ilyen modelleket felhasználó kutatások esetében fontos lenne kitérni a környezeti terhelés szempontjára is, mind a modellválasztás, mind pedig a modell használata esetében (Katona, Szeitl & Túry-Angyal, 2025).

Bár a nagy nyelvi modellek kapcsán felmerülhetnek kihívások, például az előítéletek és torzítások kezelése, ezek a technológiák mégis jelentős előrelépést hoztak a természetes nyelvfeldolgozás területén. Képesek hatékonyan azonosítani nyelvi mintázatokat, segíteni a szövegek rendszerezésében és automatizálni összetett nyelvi feladatokat. Éppen ezért érdemes megvizsgálni, hogy hogyan alkalmazhatók a szöveganalitikai annotáció területén.

2.3. A nagy nyelvi modellek alkalmazása a szöveganalitikai annotációban

A nagy nyelvi modellek egyre szélesebb körű elterjedése számos kutatót inspirált arra, hogy megvizsgálja, hogy ezek a modellek hogyan és milyen hatékonyság, valamint megbízhatóság mellett alkalmazhatók a szöveganalitikai annotációs feladatokban.

Gilardi és munkatársai (2023) négy különböző szövegkorpuszon vizsgálták a ChatGPT-3.5 Turbo zero-shot annotációs képességeit, különböző temperature beállítások mellett. Az

eredmények azt mutatták, hogy a modell szinte minden feladatban jobban teljesített, mint a laikus annotátorok, és nagyobb intercoder egyetértést mutatott, mint az emberi annotátorok közötti értékek. Hasonló eredményeket mutatott Törnberg (2023) elemzése is, amely politikai Twitter-bejegyzések pártpreferencia szerinti osztályozására fókuszált. A vizsgálatban a ChatGPT-4 nagyobb pontosságot és megbízhatóságot mutatott, mint a laikus annotátorok és a szakértők, miközben torzítás tekintetében sem teljesített rosszabbul. Ugyanakkor kiemelendő, hogy mindkét tanulmány hangsúlyozta, hogy a modellek teljesítménye érzékeny lehet a promptok megfogalmazására, valamint a zero-shot és few-shot megközelítések közötti választásra. Az említett cikkek ennek ellenére a két megközelítésből csak a zero-shot technikát vizsgálták meg. Ezzel szemben jelen szakdolgozat célja a két megközelítés pontosságának és megbízhatóságának összehasonlítása is.

Más tanulmányok azonban árnyaltabb képet mutatnak a ChatGPT annotációs képességeiről. Ostyakova és munkatársai (2023) a nyílt végű párbeszédek beszédfunkciók szerinti annotációját vizsgálva arra jutottak, hogy a megfelelő hiperparaméter-hangolás és a finomított promptok alkalmazása lehetővé teszi, hogy a modell teljesítménye elérje a laikus annotátorok szintjét, különösen strukturált, lépésenkénti kérdéssorozat alkalmazása esetén. A szerzők ugyanakkor kiemelik, hogy a modell bizonyos nehezen megkülönböztethető kategóriák esetében alulmaradt az emberi annotátorokkal szemben. Érdemes lehet tehát összevetni a modellek teljesítményét a könnyen és nehezen besorolható szövegek esetében és megvizsgálni azt, hogy miért bizonyulhatnak egyes esetek nehezebbnek a modellek számára.

Zhu és munkatársai (2023) ezzel szemben vegyes eredményeket kaptak öt különböző Twitter-alapú annotációs feladatra vonatkozóan. Míg az emócióelemzés esetében a modell elfogadható pontosságot mutatott, a gyűlöletbeszéd-azonosítás során jelentős tévesztéseket produkált. Fontos megjegyezni, hogy Zhu és munkatársai kizárólag zero-shot megközelítést alkalmaztak, és a promptokat, valamint a hiperparamétereket nem finomhangolták, ami befolyásolhatta az eredményeket.

Az LLM-ek alkalmazásának további kihívásaira hívják fel a figyelmet Ollion és munkatársai (2023), valamint Kristensen-McLachlan és munkatársai (2023) is. Kiemelten foglalkoztak a zárt forráskódú modellek használatából eredő problémákkal, különösen a reprodukálhatóság kérdésével. A szerzők rámutattak arra, hogy a zárt forráskódú modellek frissítéseivel és az újabbak megjelenésével a korábbi verziókhoz való hozzáférés megszűnhet, ami megnehezíti a

kutatások megismételhetőségét. Emellett adatvédelmi és szerzői jogi aggályok is felmerülhetnek, hiszen modellek használatakor elérhetővé tesszük az adatokat ezeknek a platformoknak.

Mindezek fényében különös jelentőséggel bírnak azok a kutatások, amelyek a nyílt forráskódú alternatívák alkalmazhatóságát vizsgálják, mivel ezek megfelelő választ nyújthatnak a closed-source modellek korábban felvetett korlátaira. Alizadeh és munkatársainak (2023) tanulmánya a nagyméretű nyílt forráskódú nyelvi modellek (pl. Llama, FLAN) teljesítményét vetette össze a ChatGPT és az MTurk annotátorok eredményeivel. Összesen négy szövegkorpuszon és tizenegy annotációs feladaton tesztelték a modelleket, eltérő temperature beállításokkal és zero-shot, illetve few-shot tanulási megközelítésekkel. Az eredmények szerint általában a ChatGPT nyújtotta a legjobb teljesítményt, de a nyílt forráskódú modellek több feladatban is felülmúlták az MTurk annotátorait. A tanulmány kiemeli, hogy bár a nyílt forráskódú modellek összességében még elmaradnak a ChatGPT teljesítményétől, több esetben megközelítik azt, miközben költséghatékonyabb, transzparensebb, reprodukálhatóbb és adatvédelmi szempontból biztonságosabb alternatívát nyújtanak. Ezen szempontokat figyelembe véve jelen kutatásom egyik célja a nyílt és zárt forráskódú modellek teljesítményének összehasonlítása, különös tekintettel az annotációs pontosságra és a reprodukálhatóságra.

A pozitív eredmények ellenére ugyanakkor a nagy nyelvi modellek megbízhatósága és pontossága egyelőre még vita tárgyát képezi. Reiss (2023) meglátása szerint a ChatGPT-3.5 zero-shot szövegannotációban mutatott megbízhatósága és konzisztenciája elmarad attól, amely a tudományos kutatásban elfogadható lenne. A szerző szerint már kisebb promptváltoztatások vagy azonos bemenetek többszöri futtatása is eltérő eredményekhez vezethet. Bár az annotációs megbízhatóság javítható a temperature érték csökkentésével és az eredmények többszöri generálásával, valamint átlagolásával, a tanulmány hangsúlyozza, hogy a ChatGPT zero-shot módú használata önálló szövegannotációs eszközként, validáció nélkül nem ajánlott. Kristensen-McLachlan és munkatársai (2023) szintén megkérdőjelezik a nagy nyelvi modellek annotációs megbízhatóságát, és arra a következtetésre jutnak, hogy egyelőre nem érdemes ezeket használni a szöveganalitikai annotációban. Emellett azt is kiemelik, hogy habár a ChatGPT-3.5 turbo és 4 jobban teljesítenek, mint az általuk vizsgált nyílt forráskódú modellek, összeségében egyik generatív nagy nyelvi modell sem ért el annyira pontos

eredményeket, mint a hagyományos felügyelt gépi tanulási modellek. Tapasztalataik szerint a promptok finomhangolása segíthet a pontosság növelésében, ennek hatása ugyanakkor inkonzisztens. Ezen eredmények tükrében a szakdolgozatomban különös figyelmet fordítok majd a promptok optimalizálására és a modellek által generált eredmények konzisztenciájának vizsgálatára.

Ollion és munkatársai (2023) továbbá arra is rámutattak, hogy a generatív modellek annotációs pontossága nagy varianciát mutat, és a modellek kiértékelésére alkalmazott metrikák és módszertani megközelítések is jelentősen eltérhetnek. Ez pedig tovább nehezíti az eredmények összehasonlíthatóságát. Ezen megállapítások figyelembevételével a szakdolgozatom során kiemelt figyelmet szentelek annak, hogy az alkalmazott módszertani keret és kiértékelési metrikák egyértelműen meghatározottak legyenek. A szerzők emellett arra is kitérnek, hogy a nagy nyelvi modellek használata még jobban hozzájárulhat ahhoz, hogy a kutatások angol nyelv központúak legyenek, hiszen ezek a modellek a kisebb nyelvek esetében sokkal kisebb tanulóhalmazzal is rendelkeznek, ami pedig rosszabb teljesítményhez vezethet.

Összességében a kutatások azt mutatják, hogy bár az LLM-ek, különösen a ChatGPT, jelentős előnyöket kínálhatnak a szöveganntációs feladatokban, elsősorban a kevesebb idő- és költségigény tekintetében (Gilardi et al., 2023; Ostyakova et al., 2023) alkalmazásuk során azonban elengedhetetlen a megfelelő validáció, a paraméterek és promptok finomhangolása és az emberi ellenőrzés a megbízhatóság és pontosság biztosítása érdekében. A nagy nyelvi modellek használatára ugyanakkor egyelőre még nincsnek bevett szokások vagy megfelelően kialakított és tesztelt módszertani keretek. A nagy nyelvi modellek alkalmazási lehetőségeinek újdonságát mutatja az is, hogy a témával kapcsolatos kutatások és szakirodalmak jelentős része egyelőre még csak preprint formában elérhető.

3. Adatok ismertetése

A nagy nyelvi modellek annotációban történő alkalmazási lehetőségét egy depresszióval kapcsolatos adatbázison vizsgálom, amely különböző online hozzászólásokat és fórumbejegyzéseket tartalmaz. A szövegkorpusz a RC2S2 kutatócsoport keretein belül jött létre. A korpuszban szereplő bejegyzések a 2016. február 15. és 2019. február 15. közötti időszakból származnak és a legnépszerűbb angol nyelvű depresszióval kapcsolatos nyilvános fórumokról származnak. A korpusz szövegei két lépésben kerültek kiszűrésre, az első lépésben

azok a thread-ek kerültek kiválasztásra, amelyekben szerepeltek a „depression” vagy „depressed” szavak. Ezt követően pedig azok a hozzászólások kerültek kiválasztásra, amelyek linkjében, topikjában vagy tartalmában megtalálható volt legalább egy a következő, depresszióval kapcsolatos kifejezésből: depression, depressed, bummer, desolation, desperation, moody, upset, gloom, hopelessness, depressant, melancholia, sorrow, unhappiness, feeling blue, depressive, depressive disorder, unipolar depression, bipolar, bipolar depression, major depression, MDD, persistent depressive disorder, PDD, cyclothymia, mood disorder, adjustment disorder, chronic fatigue syndrome, CFS, premenstrual dysphoric disorder. A duplikátumok és a 20 szónál rövidebb posztok eltávolítása és a további előfeldolgozási lépéseket követően a korpusz 67 857 szöveget tartalmazott. Ebből összesen 4485 került annotálásra. A bejegyzéseket szociológia alapszakos hallgatók címkézték fel, akik előzetesen ezzel kapcsolatos oktatáson vettek részt, a munkafolyamat során pedig a rendelkezésükre állt egy részletes útmutató is. Ebből adódóan inkább nevezhetjük őket szakértői annotátoroknak, mintsem laikusoknak. Minden szöveget két annotátor címkézett fel, egy elsődleges és ha volt ilyen, akkor egy másodlagos kategóriával. Összesen öt kategória közül választhattak: biomedikális, pszichológiai, szociológiai, irreleváns és nem besorolható. (Németh, Sik & Máté, 2020)

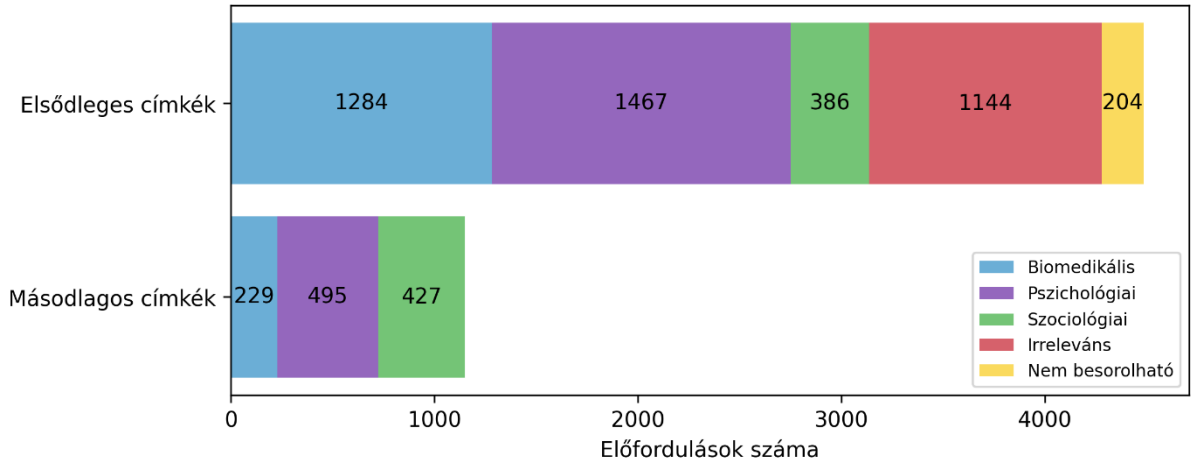
Amennyiben a két annotátor ugyanazt a címkét rendelte a szöveghez, az lett a végleges kategória. Ha csak az egyik annotátor adott érvényes címkét, míg a másik nem besorolható vagy irreleváns címkével látta el a szöveget, akkor az érvényes címke került kiválasztásra. Abban az esetben, ha az egyik annotátor egy kategóriát, míg a másik kettőt választott, és ebből kettő egyezett, akkor az egyező kategória lett kijelölve végleges címkének. Hasonlóképpen, ha a négy megadott címkéből két azonos volt, akkor az egyező kategória került kiválasztásra. Ha mindkét annotátor elsődleges és másodlagos címkét adott, és ezek sorrendje is megegyezett, akkor az elsődleges címke lett véglegesnek tekintve. Amennyiben a címkék megegyeztek, de az annotátorok eltérő sorrendben adták meg őket, egy harmadik annotátor döntése alapján került kiválasztásra a végleges címke. Végül, ha az annotátorok teljesen eltérő címkéket rendeltek a szöveghez, és egyik sem volt nem besorolható vagy irreleváns, akkor a végső döntés egy harmadik annotátor által történt meg. [4] Az összes szöveg 12,3%-ban volt szükség egy harmadik annotátor bevonására is. A bejegyzések 30%-a az irreleváns vagy nem

[4] Az adatgyűjtésről és feldolgozásról bővebben: https://ijqm.rc2s2.eu/data/

besorolható kategóriába tartozott (Németh, Sik & Máté, 2020; Németh et al., 2022). Illetve a végleges címkéket nézve a posztok nagyjából 25%-a került besorolásra egy második kategóriába is.

Az elsődleges címkéknél a pszichológiai kategória fordul elő leggyakrabban, ezt követi a biomedikális, ezzel szemben a szociológiai címkével ellátott szövegek sokkal kevésbé gyakoriak, tehát egy egyenlőtlen címkeeloszlással van dolgunk. Illetve az is látszik, hogy az irreleváns kategória sokkal jellemzőbb, mint az, hogy egy szöveg a depresszióval foglalkozik, de nem egyértelműen besorolható a három kategória egyikébe. A másodlagos címkék között a pszichológiai és a szociológiai kategóriák előfordulása dominál, míg a biomedikális másodlagos címkék száma lényegesen alacsonyabb. Külön érdekes, hogy a szociológiai kategória többször szerepel a másodlagos címkék között, mint az elsődleges címkéknél, ami arra enged következtetni, hogy ez a megközelítés általában nem fő keretezésként jelenik meg, hanem jellemzőbb az, hogy valamilyen másik kategóriával együtt fordul elő a szövegben.

1. Ábra: A különböző kategóriák gyakorisága az elsődleges és másodlagos címkék között



Az általam felhasznált adatokból számos tanulmány és szakdolgozat született, amelyek közül néhányat röviden ismertetek. Németh, Sik és Máté (2020) részletesen kifejtik, hogy a keretezési kategóriákban történő besorolás még az emberi annotátorok számára is nehézségekkel járt, az annotátorok közötti egyetértés pedig még az annotáció pilot szakaszának meghosszabbítása és az utasítások folyamatos finomítása mellett is elmaradt a vártól. A szerzők arra is kitérnek, hogy az annotált adatokon alapúló gépi tanulási módszerek a biomedikális és a pszichológiai kategóriákban jól teljesítenek, ezzel szemben a szociológia kategória beazonosításában nehézségekbe ütköznek, amely arra vezethető vissza, hogy a

depresszió szociológiai keretezése esetében nincsenek olyan mértékben jelen a bevett szakszavak, ezáltal pedig nagyobb hangsúly helyeződik a hermeneutikai jelentésalkotási folyamatra, amelyet azonban a gépi tanulási módszerek nem képesek reprodukálni, hiszen ezek az eszközök elsősorban a különböző nyelvi mintázatok felismerésében jeleskednek. Érdekes kérdés, hogy vajon a nagy nyelvi modellek esetében is fennáll-e majd ez a limitáció vagy ezek az eszközök képesek-e egy emberi jelentéstulajdonító gondolkodáshoz hasonló értelmezése.

Egy másik cikkükben Németh és munkatársai (2022) ugyanezen klasszifikációs problémát számos gépi tanulási eljárással vizsgálták meg. A modellek tanítása során azt is figyelembe vették, hogy egy szöveg könnyen besorolhatónak számított-e (a két annotátor egyetértett-e az elsődleges címkében). Az eredmények alapján az a modell teljesített a legjobban, amely könnyen besorolható szövegeken lett tanítva. Ezek alapján az annotátorok számára nehéz esetek a gépnek is nehezek voltak. Ha kizárólag ezeken történt a modell tanítása az rosszabb eredményekre vezetett. Azonban, ha a nehéz eseteket vegyítették a könnyű esetekkel, ez a negatív hatás csökkent. A kizárólag vagy részben könnyű adatokon történő tanítás azonban nem rontotta el a nehéz esetek besorolását, amely a szerzők szerint arra enged következtetni, hogy ezek a nehezen besorolható szövegek hasonlítanak a könnyű esetekre, de a modell nem tudta felismerni ezt a hasonlóságot, amikor kizárólag nehéz eseteken tanult. A kutatás arra is rávilágít, hogy a nehéz esetek többsége szociológiai keretezésű volt, míg a könnyebb esetek inkább biomedikálisak. A könnyebb szövegek formálisabbak és pontosabbak voltak, míg a nehezen osztályozható szövegek informálisabbak és közelebb álltak a beszélt nyelvhez. A cikk egy fontos megállapítása, hogy a hasonló vizsgálatokban érdemes részletesebben foglalkozni a nem egyértelmű esetekkel is. Ezt a megállapítást a nagy nyelvi modellek teljesítményének értékelése során én is igyekszem majd figyelembe venni a vizsgálatomban.

A dolgozatom szempontjából érdekes még Kerekes Norbert (2020) szakdolgozata, aki ugyanezen adatokon vizsgálta a multi-label szövegklasszifikáció kérdéskörét különböző gépi tanulási módszerekkel. A dolgozat szerzője a korábban ismertetett cikkekkel összhangban szintén arra jutott, hogy a tanuló algoritmusok a biomedikális keretezést ismerték fel a legkönnyebben. Emellett az eredmények azt mutatták, hogy az olyan modellek, amelyek figyelembe veszik a címkék közti kapcsolatokat jobb teljesítményt nyújtottak, különösen a szociológiai keretezés azonosításában. Kerekes megállapítása alapján az adatbázisra alacsony

címkesűrűség volt jellemző, ami azt eredményezte, hogy a modellek hajlamosak voltak „nullás kimeneteket” jósolni, vagyis nem rendeltek címkéket a bejegyzésekhez.

4. Módszertan

4.1. Választott nagy nyelvi modellek

A vizsgálatom során két nagy nyelvi modell annotációs teljesítményét hasonlítottam össze. Ezek közül az egyik az OpenAI GPT-4o mini modellje, ez egy 2024. júliusában megjelent zárt forráskódú multimodiális modell. A modell dokumentációja[5] alapján a GPT-4o mini egy kisebb modell, ennek ellenére azonban jobb teljesítményt nyújt, mint a korábbi GPT-3,5 turbo. Az OpenAI belső tesztjei szerint a GPT-4o mini több fontos területen is jobb teljesítményt nyújt, mint a GPT-3.5 turbo. Az általános nyelvi tudást mérő MMLU teszten erősebb, ami azt jelzi, hogy jobban ért szövegek tartalmához. Az olvasott szöveg értelmezésében (DROP) és a matematikai feladatok megoldásában (MATH, MGSM) szintén megbízhatóbb. Kódgenerálásban (HumanEval) is jobban teljesít, vagyis összetettebb logikai feladatokat is hatékonyabban old meg. Emellett a következtetést és problémamegoldást igénylő feladatokban, például a GPQA és MathVista benchmarkokon is magabiztosabb, mint a 3.5-ös verzió. A modell fő előnye a költség-és energiahatékony működése, ami miatt különösen alkalmas lehet abban az esetben, ha nagy mennyiségű szöveges adatot szeretnénk feldolgozni és ezzel együtt különböző megközelítéseket tesztelni. A GPT-4o mini modell 128 ezer token méretű kontextusablakkal[6] rendelkezik, ezzel lehetővé téve a hosszabb szövegek és összetettebb annotációs feladatok kezelését is.

Az általam vizsgált másik nagy nyelvi modell a Meta által fejlesztett Llama 3.3, amelynek 70 milliárd paraméterrel rendelkező verziója 2024. decemberében vált elérhetővé. A modell teljes mértékben nyílt forráskódú, ezáltal átláthatóbb és reprodukálhatóbb kutatási eredményeket biztosíthat, emellett lokálisan futtatva pedig költséghatékonyabb is. A modell dokumentációja[7] és a Meta mérési eredményei alapján a Llama 3.3 70B több benchmarkon felülmúlja a korábbi,

[5] Bővebben a GPT-4o mini modellről: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

[6] A kontextusablak a nyelvi modellek azon kapacitását jelöli, hogy egyszerre hány tokennyi információt képesek figyelembe venni a válaszaik során. Minél nagyobb ez az ablak, annál hosszabb szövegek értelmezése és feldolgozása válik lehetővé anélkül, hogy a korábbi szövegrészek kiesnének a modell memóriájából.

[7] Bővebben a Llama 3.3 70B modellről: https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md

azonos paraméterszámú Llama 3.1 70B Instruct verziót. Az általános nyelvi megértést mérő teszteken (MMLU, MMLU Pro) megbízhatóan teljesít, és az utasítások pontos követésében (IFEval) is jelentősen javult. Kódolási feladatokban (HumanEval, MBPP) jobban szerepel, vagyis hatékonyabban old meg programozási problémákat. A matematikai feladatokban (MATH) és többnyelvű szövegértésben (MGSM) is előrelépés történt, ami a modell fejlettebb logikai és nyelvi képességeire utal. A modell a GPT-4o minihez hasonlóan 128 ezer token méretű kontextusablakkal rendelkezik.

A modellek futtatása mindkét esetben API-n keresztül történt. A GPT-4o mini esetében erre más lehetőség nem is lett volna, hiszen nagy mennyiségű adat feldolgozásáról van szó, így a Chat felület használata nem lehetséges, a modell pedig lokálisan nem futtatható csak az OpenAI API rendszerén keresztül. Ezzel szemben a Llama 3.3 70B esetében lehetséges a modell lokális futtatása, azonban megfelelő hardver hiányában erre nem volt lehetőségem, így a Groq[8] API rendszerén keresztül végeztem el a futtatást. Ez a felállás abból a szempontból is optimális, hogy a Groq rendszere az OpenAI API behívásához hasonlóan működik, így a kódokon csak minimális módosításokat szükséges elvégezni, a két rendszer jól átjárható. A modellek API-n keresztül történő futtatását és az eredményeik kiértékelését Pythonban végeztem el.

4.2. Címkék generálása

A GPT-4o mini és a Llama 3.3 modellek már előzetesen betanításra és részleges finomhangolásra kerültek, azonban mindkét modell esetében adott a további finomhangolás lehetősége is. Ugyanakkor a szakdolgozatomban ezzel a lehetőséggel nem éltem, hanem a hangsúlyt inkább arra helyeztem, hogy részletesen megvizsgáljam a különböző promptolási technikák és temperature beállítások hatását a modellek annotációs teljesítményére. Ezáltal nem klasszikus értelemben vett modell-tanítás történik, hanem a promptok és paraméterek optimalizálása. Ezzel a megközelítéssel jelentősen csökkenthető a túlillesztés kockázata is.

Ennek ellenére a rendelkezésre álló 4485 annotált szöveget tanító- és teszthalmazra osztottam 20/80 arányban. A „tanítóhalmaz" célja ebben az esetben a promptok és a temperature paraméter optimális beállításának meghatározása, míg a teszthalmaz segítségével értékelem a tanítóhalmazon legjobban teljesítő modellek eredményeit, így

[8] A Groq egy API szolgáltató, amely több különböző nyílt forráskódú modell futtatását teszi elérhetővé. https://console.groq.com/docs/overview

elkerülhető például a promptok túlspecifikálása. Tekintettel arra, hogy az adatokban szereplő kategóriák aránya jelentősen eltér, a tanító- és teszthalmazokat rétegzett mintavétellel hoztam létre úgy, hogy mindkét halmaz tükrözze a teljes adatbázis kategóriaeloszlását. Hagyományosan általában a tanítóhalmaz mérete a nagyobb, azonban jelen kutatás esetében több okból is úgy döntöttem, hogy eltérek ettől a felállástól. Egyrészt, mint ahogy azt már említettem, szigorúan véve tényleges modelltanítás nem történik, így nincs szükség arra sem, hogy a használt modellek minél több adaton tudjanak betanulni. Másrészről az adatok ilyen arányú felbontása a hatékonysági szempontból is ideálisabb. Azzal, hogy az adatok nagyjából ötödén optimalizálom a különböző felállásokat, csökken a modellek futtatási ideje, a futtatások költsége, illetve a felhasznált energia is.

A nagy nyelvi modellekkel történő annotáció esetében a puha és a kemény klasszifikációs megközelítések eredményességét egyaránt megvizsgáltam. A kemény klasszifikáció során a modell bináris döntést hoz, hogy egy adott dokumentum tartozik-e a kategóriához vagy sem. A puha klasszifikáció esetében a modell minden kategóriához egy numerikus értéket rendel, amely a dokumentum adott kategóriába tartozásának valószínűségét vagy relevanciáját jelzi. Ezt követően egy meghatározott küszöbérték alapján dől el, hogy a dokumentum mely kategóriákba kerül besorolásra (Wahba, 2002). Mindkét módszer vizsgálata lehetőséget nyújtott arra, hogy részletesebben értékeljem a modellek kategorizációs teljesítményét és megértsem a klasszifikációval kapcsolatos esetleges bizonytalanságokat.

A modellek annotációs teljesítményét különböző promptolási megközelítésekkel vizsgáltam: zero-shot és few-shot technikákkal. A zero-shot megközelítés egy rövidebb esetében a modellek számára kizárólag a kategóriák neveit és a feladat részletes leírását nyújtottam. A zero-shot módszer egy másik változatában a modellek ezen kívül a feladat és a kategóriák részletesebb leírását is megkapták, így segítve azok pontosabb értelmezését. Ezzel szemben a few-shot megközelítés során minden kategóriához példaszövegeket is mellékeltem, amelyek célja a kategóriák pontosabb értelmezésének és az annotációs teljesítmény javításának elősegítése volt.

A modellek teljesítményének vizsgálatakor még egy szempontot figyelembe vettem, ami nem más, mint a temperature paraméter értéke. A paraméter segítségével tulajdonképpen a modell „kreativitását” szabályozhatjuk. Alacsonyabb temperature értékek esetén a válaszok konzisztensebbek és determinisztikusabbak, míg magasabb értékek esetén kreatívabb és

változatosabb válaszokat kapunk. Ennek oka, hogy alacsonyabb értékek esetén a modell kisebb valószínűséggel választ ritka vagy nem valószínű szavakat a generálás során, míg nagyobb értékek esetén nagyobb valószínűséggel generál ilyen szavakat (Jurafsky & Martin, 2025). Vizsgálatom során négy különböző temperature értékkel dolgoztam: 0 (determinisztikusabb), 0,2 (alacsony variabilitás), 0,4 (mérsékelt variabilitás) és 0,6 (nagyobb variabilitás).

A generálás szabályozására létezik egy másik gyakran használt paraméter is, az úgynevezett top_p vagy nucleus sampling, amely a tokenek azon legkisebb halmazát veszi figyelembe, amelynek kumulatív valószínűsége eléri a megadott p értéket (Holtzman et al., 2020). Alapértelmezett értéke egy, ami azt jelenti, hogy a teljes valószínűségi eloszlásból választhat, vagyis több lehetséges szóból, ezáltal változatosabb eredményeket adva. Ezzel szemben az alacsonyabb értékek esetében kevesebb szó közül választhat a modell a generálás során. Mivel vizsgálatom során kizárólag a temperature változtatására koncentráltam, a top_p értékét minden esetben az alapértelmezett egyes értéken hagytam.

A két klasszifikációs megközelítés, három promptolási technika és négy temperature érték kombinációja modellenként összesen 24 különböző esetet foglal magába, amelyek az 1. táblázatban láthatók. Ebből a 24-ből végül a GPT-4o mini és a Llama 3.3 estében is két-két megközelítést választottam, tehát részletesebben összesen négy modell elemzésére koncentráltam, ennek okaira a modellszelekciós alfejezetben részletesebben is kitérek.

1. táblázat: A tanítóhalmazon megvizsgált modell elrendezések

Klasszifikációs megközelítés	Promptolási technika	Temperature értékek
Kemény klasszifikáció	Zero-shot (csak kategórianevek)	0; 0,2; 0,4; 0,6
	Zero-shot (kategóriák részletes leírása)	0; 0,2; 0,4; 0,6
	Few-shot (kategóriák + példák)	0; 0,2; 0,4; 0,6
Puha klasszifikáció	Zero-shot (csak kategórianevek)	0; 0,2; 0,4; 0,6
	Zero-shot (kategóriák részletes leírása)	0; 0,2; 0,4; 0,6
	Few-shot (kategóriák + példák)	0; 0,2; 0,4; 0,6

A promptok megalkotásához többféle forrást is felhasználtam annak érdekében, hogy a nagy nyelvi modellek számára egyértelmű és a feladathoz illeszkedő utasításokat biztosítsak. Az egyik legfontosabb kiindulási pont az az annotátorok számára készített részletes útmutató volt, amely a korábbi, emberi címkézési folyamat során szolgált segítségül. Ez az útmutató

egyrészt felsorolja az egyes kategóriákat (biomedikális, pszichológiai, szociológiai, irreleváns és nem besorolható), emellett pedig azok részletes jellemzőit, valamint számos példát is tartalmaz magyarázatokkal kiegészítve. Ezen kívül támaszkodtam Németh, Sik és Máté (2020), illetve Németh és munkatársai (2022) tanulmányaira is, amelyekben részletesen bemutatják az annotációs folyamat módszertanát, a keretezési kategóriák kialakításának elméleti hátterét, valamint azt is, hogy mit jelent pontosan a depresszió keretezése.

A promptok kialakítása során végül az annotált szövegek egy részét is aktívan felhasználtam. Ezeket az eseteket példaként beépítettem a few-shot promptokba annak érdekében, hogy a modell számára kontextusban, konkrét példákon keresztül legyenek érzékelhetők az egyes kategóriák közötti különbségek és hasonlóságok. Így a promptok nemcsak elméleti definíciókat, hanem gyakorlati példákat is tartalmaztak, amely segített a pontosság növelésében. Természetesen a példaként felhasznált posztok a tanító- és a teszthalmazból egyaránt kikerültek. Minden kategóriához 2-2 példát adtam meg a modelleknek, a biomedikális, a pszichológiai és szociológiai elsődleges címkékkel rendelkező szövegek esetében ezek közül az egyik példa minden esetben olyan volt, amelyhez másodlagos címke is tartozott. Ezzel ösztönözve a modellt arra, hogy ha van másodlagos keretezés, akkor arra is adjon besorolást.

A kemény klasszifikációs esetek során arra kértem a modellt, hogy minden egyes szöveg esetében kizárólag egy, vagy legfeljebb két kategóriát jelöljön meg, mégpedig elsődleges, illetve ha indokoltnak látja, másodlagos címkeként. Ebben az esetben a prompt arra utasította a modellt, hogy válasszon ki egy kategóriát az előre meghatározott ötből (biomedikális, pszichológiai, szociológiai, irreleváns, nem besorolható), és ha szükséges, adjon meg egy másodlagos címkét is. A másodlagos címkék esetében az volt a kérés, hogy csak a biomedikális, pszichológiai és szociológiai címkéket használja, hiszen ha már adott egy „érvényes” választ elsődleges címkeként, akkor nem logikus, hogy a másodlagos irreleváns vagy nem besorolható legyen. Ez a megközelítés azt modellezte, ahogyan a humán annotátorok is dolgoztak a manuális címkézés során. A két kategória meghatározása mellett még egy konfidencia pontszámot vártam el a modelltől, amelyben azt kellett megmondania, hogy egy 0-100-ig tartó skálán mennyire biztos a klasszifikációjával.

A puha klasszifikációs megközelítés során arra kértem a modellt, hogy ne egy vagy két konkrét kategóriát válasszon ki az adott szöveghez, hanem mind az öt lehetséges kategóriához

rendeljen valószínűségi értékeket 0 és 1 között. Ez a megközelítés lehetővé tette, hogy árnyaltabb képet kapjak arról, hogy a modell mennyire tart jellemzőnek egy adott keretezést az adott bejegyzésre vonatkozóan. Illetve ennél a megközelítésnél is megkértem a modellt arra, hogy adjon egy konfidencia pontszámot.

A promptok megírása egy többlépcsős folyamat volt. A kezdeti változatokat többször is módosítottam a modellek válaszai alapján, mindig annak érdekében, hogy azok jobban kövessék a feladat leírását és a megadott válaszformátumot. A módosítások során új utasításokat építettem be a promptokba, amelyek segítették a modelleket abban, hogy pontosabb, következetesebb válaszokat adjanak.

A promptok használata során több olyan probléma is felmerült, amelyek a modellek válaszainak pontosságát, értelmezhetőségét vagy egységességét rontották. Ezek a nehézségek főként a válaszformátum betartásával, a másodlagos címkék kezelésével, a valószínűségi értékek megadásával, illetve a bizonytalan esetek kezelésével kapcsolatosak voltak. A problémák felmerülése után a promptokat célzott utasításokkal egészítettem ki, annak érdekében, hogy a modellek a feladatot az elvárásoknak megfelelően hajtsák végre.

Az egyik lényegi probléma az volt, hogy a modellek nem tartották be az elvárt válaszformátumot és szöveges magyarázatokat is adtak válaszként, vagy nem minden szükséges adatot adtak meg. Ennek kezelésére kifejezetten hangsúlyoztam a promptban, hogy az adott példában szereplő formátumot kövessék, és semmi mást ne írjanak: *„Follow this output example exactly: '0.65 0.25 0.08 0.01 0.01 85'. Do not write anything else.”*[9] Hasonló utasítást adtam a kemény klasszifikációs esetekre is, ahol három értéket vártam: elsődleges címke, másodlagos címke (vagy 999, ha nem volt másodlagos kategória), illetve egy konfidencia pontszám. A promptnak ebben az esetben arra is ki kellett térnie, hogy minden egyes értékre nyújtson választ a modell, mert néhány alkalommal csak részleges eredményt adott vissza: *„You must always provide all three values. Do not write anything else, besides the three numbers.”*[10]

[9] A prompt magyar fordítása: „Kövesd pontosan ezt a kimeneti példát: '0.65 0.25 0.08 0.01 0.01 85'. Ne írj semmi mást.”

[10] A prompt magyar fordítása: „Mindig meg kell adnod mindhárom értéket. Ne írj semmi mást a három számon kívül.”

A kemény klasszifikációnál a részletesebb zero-shot és few-shot promptoknál szintén gyakori hiba volt, hogy a modell másodlagos címkeként a „nem besorolható” kategóriát (4) adta meg, amely az annotációs logika szerint nem értelmezhető másodlagos keretezésként. Ennek elkerülése érdekében külön kiemeltem a modelleknek, hogy ezt ne használják: *„You should never use four as a secondary label”*[11]. Emellett pontosan meghatároztam, milyen kategóriák adhatók meg másodlagos címkeként (1–3 vagy 999).

A puha klasszifikációs megközelítésnél további problémát jelentett, hogy a modell hajlamos volt több kategóriához is azonos valószínűségi értéket rendelni (pl. három kategóriára egyaránt 0,33-at), ami nehezítette az elsődleges és másodlagos kategóriák meghatározását. Ezt a problémát olyan utasításokkal kezeltem, amelyek előírták, hogy a nem nulla értékek legyenek különbözőek: *„Ensure that each non-zero value is slightly different from the others”*[12], illetve *„Only assign the same value to multiple categories if that value is 0.00.”*[13]

Végül a puha klasszifikáció során olyan esetekkel is találkoztam, amikor a modell bizonytalan volt a döntésben, és ennek következtében minden kategóriára 0-át adott meg, ami a gyakorlatban nem hasznosítható. Az ilyen válaszok elkerülése érdekében a promptba bekerült az a kérés is, hogy ha bizonytalan a válaszában, akkor is próbáljon meg adni valószínűségi értéket a legjobban odaillő kategóriának: *„If you're uncertain, estimate the most plausible category rather than assigning all zeros.”*[14]

Összességében a promptok finomhangolása során nemcsak a tartalmi pontosságra törekedtem, hanem arra is, hogy a válaszok technikailag jól feldolgozhatók és egységesek legyenek. A fenti problémák jól mutatják, hogy a nagy nyelvi modellekkel való munka során sok esetben nehéz pontos, valamint tartalmilag és formailag helyes válaszokat kapni.

Első körben a promptokat a GPT-4o mini modellen teszteltem és finomítottam, majd ezeket a változatokat kipróbáltam a Llama 3.3-as modellen is. Ha a két modell különbözően viselkedett, a promptot újra átírtam, majd módosításokat mindkét rendszeren újra megnéztem. Felmerült annak lehetősége is, hogy a két modellhez külön-külön optimalizáljam a promptokat, de ez megnehezítette volna az eredmények összehasonlítását. Ezért inkább

[11] A prompt magyar fordítása: „Soha ne használd a négyest másodlagos címkeként.”
[12] A prompt magyar fordítása: „Gondoskodj róla, hogy minden nem nulla érték kissé különbözzön a többitől.”
[13] A prompt magyar fordítása: „Csak akkor add ugyanazt az értéket több kategóriának, ha az az érték 0.00.”
[14] A prompt magyar fordítása: „Ha bizonytalan vagy, inkább becsüld meg a legvalószínűbb kategóriát, minthogy mindenre nullát adj.”

olyan megoldásokra törekedtem, amelyek mindkét modellnél jól működtek, így biztosítva az összehasonlíthatóságot és az egyenlő feltételeket. A különböző modelleknél használt végső promptok az 1. számú függelékben találhatóak meg.

4.3. Pontosság kiértékelése

A kutatásomban a modellek kiértékeléséhez számos statisztikai mérőszámot használtam, amelyek segítenek megállapítani azok pontosságát, illetve azt, hogy milyen mértékben egyeznek az emberi annotációval. Az alkalmazott metrikák közé tartoznak a precision, recall, F1-score, valamint a Cohen-féle kappa értékek. A precision, recall és F1-score olyan mutatók, amelyeket rendszeresen használnak a gépi tanulási modellek eredményeinek kiértékeléséhez. Jelen esetben azonban egy multi-class és multi-label problémával állunk szemben. Ahogy arra Németh és munkatársai (2022) is kitérnek az említett mutatók azonban általában két címkére optimalizáltak, több kategória esetében pedig átlagolnak az összes lehetséges bináris címkepárra (például biomedikális vs. egyéb), ez viszont egyenlőtlen címkeeloszlásnál nem ideális megoldás. A szerzők megoldásként a Cohen-féle kappa alkalmazását javasolják, amelynek egy lényeges pozitívuma, hogy figyelembe veszi a véletlen egyezés valószínűségét is. Ezen javaslatot követve a szakdolgozatomban egyaránt kitérek a Cohen-féle kappa és a hagyományos mutatók eredményeire is. Ezen metrikák használata során a teljes modell teljesítményének globális értékelése mellett részletes kategóriánkénti elemzést is végeztem. Figyelembe véve az egyenlőtlen címke eloszlást a precision, recall és F1-score esetében a modell globális teljesítményére vett mutatót súlyozott átlag segítségével határoztam meg.

A precision azt mutatja meg, hogy a modell által pozitívnak ítélt elemek közül hány volt valóban pozitív.

$$Precision = \frac{TP}{TP + FP}$$

ahol: TP (true positive) a helyesen pozitívnak ítélt elemek száma, FP (false positive) pedig a helytelenül pozitívnak ítélt elemek száma.

A recall mutató a modell érzékenységét méri. Azt jelzi, hogy a valóban pozitív elemek közül mennyit sikerült helyesen besorolnia a modellnek.

$$Recall = \frac{TP}{TP + FN}$$

ahol: FN (false negative) a tévesen negatívnak besorolt elemek száma.

Az F1-score (F1) a precision és recall harmonikus átlaga, amely egy mérőszámban foglalja össze a modell pontosságát és érzékenységét.

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall}$$

A Cohen-féle kappát (Cohen, 1960) gyakran az annotátorok közötti egyetértés vizsgálatára használják, de jelen esetben azt mutatja meg, hogy mekkora az egyezés szintje a modell és az emberi annotáció között, illetve hogy mennyivel teljesít jobban a modell egy olyan véletlen osztályozó eljáráshoz képest, amely minden kategóriát a relatív gyakoriságuk szerint választana ki.

$$\kappa = \frac{p - p_e}{1 - p_e}$$

ahol p a megfigyelt egyezés aránya, p_e pedig a véletlen egyezés esetén várható érték.

Ha a megfigyelt egyezés $p = 1$, azaz tökéletes egyezés van, akkor $\kappa = 1$. Ha a modell véletlenszerűen választja a címkéket, akkor $p = p_e$, és ekkor $\kappa = 0$ (Cohen, 1960; Németh et al., 2022).

Németh, Sik és Máté (2020), illetve Németh és munkatársai (2022) megközelítését követve minden pontossági mutatót két módon számoltam ki. A konzervatív nézőpontot alkalmazva egy szöveg akkor tekinthető helyesen besoroltnak, ha a modell által adott elsődleges címke megegyezik a valódi elsődleges címkével. Ekkor a másodlagos címkék nincsenek figyelembe véve, ez azonban egy túlságosan szigorú eredményre vezet, amely a részleges egyezéseket nem jutalmazza. Flor és munkatársai (2016) megközelítéséhez hasonlóan Németh, Sik és Máté (2020), illetve Németh és munkatársai (2022) liberális módon is kiértékelték a modellek eredményeit. A liberális megközelítést alkalmazva a részleges egyezéseket is pontos klasszifikációként vehetjük figyelembe, ebben az esetben tehát akkor is pontosnak tekinthető egy besorolás, ha: a modell elsődleges címkéje megegyezik a valódi elsődleges címkével; ha a modell által adott elsődleges címke nem egyezik a valódi elsődlegessel, de egyezik a valódi másodlagossal; illetve ha a valódi elsődleges címke nem egyezik a modell által adott elsődleges címkével, de egyezik a modell által adott másodlagos címkével. Ahogy arra Németh és munkatársai is kitérnek a két megközelítés együttes értelmezése lehet a legjobb megközelítés,

ugyanis a liberális mutatók esetleg túl engedékenyek lehetnek, a konzervatív módszer viszont a másodlagos címkék figyelmen kívül hagyásával túl szigorú eredményre vezet.

A szakdolgozatomban azt is megvizsgáltam, hogy a modellek teljesítménye hogyan alakul a könnyen és a nehezen besorolható eseteknél, ezekben a csoportokban külön-külön is kiszámítottam a fenti metrikákat. Ezzel a célom annak megvizsgálása volt, hogy vajon a modellek számára is azok a szövegek számítanak-e könnyebbnek, amelyek esetében az annotátorok is egyszerűbben jutottak egységes döntésre. Németh és munkatársaihoz (2022) hasonlóan könnyűnek azokat az eseteket vettem, ahol az annotátorok egyetértettek az elsődleges címkében, míg nehezeknek azokat, ahol mindez nem állt fenn.

A modellek által megadott konfidencia pontszámokat szintén felhasználtam a pontosság vizsgálatában. Ebben az esetben arra voltam kíváncsi, hogy ezek az értékek mennyire tükrözik a modellek bizonytalanságát, illetve hogy a magasabb konfidencia pontszám nagyobb pontossággal jár-e. Bár a skála elvileg 0 és 100 közötti értékeket vehet fel, a gyakorlatban azt tapasztaltam, hogy a modellek elsősorban a két végletet használják, vagy 0-t adnak meg, vagy pedig 70 és 100 közötti értékeket, ráadásul jellemzően ezeket is ötösével lépegetve. A köztes tartományok ritkán jelentek meg az eredményekben, ezen okokból kifolyólag a változó eredetileg tervezett folytonos mivolta erősen sérült.

Ezért ahelyett, hogy a nyers pontszámokat használtam volna, a további elemzéshez három kategóriába soroltam a válaszokat a konfidencia pontszámok alapján. Ez lehetővé tette, hogy a modellek teljesítményét összehasonlítsam a különböző csoportokban. Mindegyik kategórián belül külön-külön kiszámítottam a pontossági mutatókat (precision, recall, F1-score, Cohen-féle kappa), így láthatóvá vált, hogy a nagyobb konfidencia pontszámmal rendelkező csoportokban magasabb-e a modellek pontossága is. A vizsgálat célja az volt, hogy választ kapjak arra, vajon érdemes-e figyelembe venni a modellek által megadott konfidencia pontszámot, hiszen ha ezek indikálják a pontos besorolást, akkor az annotációs folyamat során érdemes lehet csak azokat az eseteket ellenőrizni, ahol alacsony a modellek által megadott érték.

4.4. Megbízhatóság kiértékelése

A dolgozatomban nemcsak a modellek pontosságát, hanem azok megbízhatóságát is igyekeztem megvizsgálni. A megbízhatóság alatt itt azt értem, hogy egy adott modell mennyire

következetesen hozza ugyanazt a döntést többszöri futtatás esetén, illetve hogy a különböző modellek mennyire jutnak hasonló eredményekre ugyanazon szövegek esetében.

A vizsgálathoz ebben az esetben is a Cohen-féle kappa mutatót használtam, mivel ez az egyik legelterjedtebb mérőszám az annotátorok közötti egyetértés vizsgálatára. A kappa figyelembe veszi a véletlenszerű egyezés valószínűségét is, így árnyaltabb képet ad, mint az egyszerű egyezési arány, és lehetőséget biztosít arra is, hogy az eredményeket közvetlenül összehasonlítsam a korábbi, ugyanezen adatbázison végzett kutatásokkal (Németh, Sik & Máté, 2020; Németh et al., 2022).

Mivel a Cohen-féle kappa csak két annotátor, vagyis jelen esetben két modellkimenet közötti egyezést képes mérni, a végső modellek esetében a vizsgálatot páronkénti összehasonlítás formájában végeztem el. Az egyes modelleket három alkalommal futtattam le azonos prompttal és beállításokkal, majd az egyes futáspárokhoz (1–2, 1–3, 2–3) külön-külön kiszámítottam a kappa értékeket, végül ezek átlagolásával hoztam létre egy összesített megbízhatósági mutatót.

Emellett megvizsgáltam azt is, hogy különböző modellek, pontosabban GPT-4o mini és a Llama 3.3 70B milyen mértékben jutottak azonos eredményre, amikor ugyanazt a szövegkorpuszt ugyanazzal a prompttal dolgozták fel. Ennek célja az volt, hogy felmérjem, mennyire hoznak egységes döntést eltérő modellek, ha azonos inputot kapnak.

Ezen felül azt is megvizsgáltam, hogy ugyanazon a modellen belül az eltérő promptolási stratégiák mennyire jutnak ugyanarra az eredményre. Ez a vizsgálat lehetőséget adott annak feltérképezésére, hogy mennyire stabil a modell döntéshozatala a promptok változásának hatására, illetve hogy milyen mértékben befolyásolja a válaszokat az utasítás formája, részletessége vagy példák jelenléte. A célom az volt, hogy felmérjem, mennyire tekinthetők robusztusnak az egyes promptolási megközelítések.

4.5. Modellszelekció

A modellek pontosságának és megbízhatóságának részletes értékelését a teszthalmazon csak azokban az esetekben végeztem el, amikor az adott modell–prompt kombináció a tanítóhalmazon már előzetesen is jól teljesített (a tanítóhalmazon futtatott modellek kappa eredményei a 2. számú függelékben megtalálhatók). A végső, részletes összehasonlításra így két-két kiválasztott modell került be mind a Llama 3.3 70B, mind a GPT-4o mini esetében.

A modellszelekció során több tényezőt is figyelembe vettem. Először is, minden promptolási stratégiát négy különböző temperature beállítással is kipróbáltam (0, 0,2; 0,4; 0;6), azonban a tapasztalatok azt mutatták, hogy ez a paraméter nem gyakorolt jelentős hatást a modellek teljesítményére. Nem fordult elő olyan eset, ahol a magasabb temperature értékek számottevően pontosabb eredményhez vezettek volna. Ennek egyik lehetséges oka az, hogy a feladat során a modelleknek kizárólag számokat kellett visszaadniuk meghatározott formátumban, és nem kellett szöveges választ vagy összetett nyelvi szerkezetet generálniuk. A temperature paraméter hatása így valószínűleg jóval kisebb, mint például egy kreatív szövegalkotási feladatnál lenne. Ezen kívül azt is tapasztaltam, hogy a magasabb temperature értékek mellett a modellek gyakrabban tértek el az utasítások pontos követésétől. Emellett mivel a determinisztikusabb működés előnyösebb az annotációs konzisztencia szempontjából, és ez nem járt együtt a pontosság romlásával, a végső összehasonlításban kizárólag a 0-ás temperature értékkel futtatott modelleket vettem figyelembe.

Ezen kívül a zero-shot promptolási technikán belül két különböző megközelítést is kipróbáltam. Az egyik esetben a modell csak a feladat rövid utasítását és a választható kategóriák listáját kapta meg, míg a másik változatban részletes magyarázat is szerepelt arról, hogy mit jelentenek az egyes kategóriák. A kísérletek alapján, nem túl meglepő módon ez utóbbi verzió jobb teljesítményt nyújtott mind a GPT-4o mini, mind a Llama 3.3 esetében, ezért a további tesztelést és értékelést kizárólag erre a változatra végeztem el.

Továbbá, a kezdeti szakaszban a puha klasszifikációs megközelítést is megvizsgáltam, a végső elemzés során végül mégis kizárólag csak a kemény klasszifikációs eredményekkel dolgoztam tovább. Ennek több oka is volt, amelyek elsősorban a puha címkézés technikai és értelmezési nehézségeiből adódtak.

A puha klasszifikáció során az elsődleges címkét minden esetben az a kategória jelentette, amelyre a modell a legnagyobb valószínűségi értéket adta. Mivel a kategóriák között az irreleváns és a nem besorolható is ott volt, így nem volt szükség egy döntési hattára vagy optimalizálásra az elsődleges címkék esetén, hiszen ha érvényes kategória nem volt, akkor a legnagyobb valószínűséget ezek közül kapta valamelyik. Azonban az utasítások ellenére több alkalommal is előfordult, hogy a modellek több kategóriához is pontosan ugyanakkora valószínűséget rendeltek. Ezekben az esetekben nem volt egyértelmű, melyik kategóriát tekintsük elsődlegesnek, ebből adódóan vagy véletlenszerűen kellett volna választani, vagy

további újrafuttatással egyértelműsítő választ kérni a modellektől. Az instrukciók ellenére a Llama modellnél emellett olyan esetek is előfordultak, amikor a modell minden kategóriára 0-ás értéket adott, vagyis gyakorlatilag nem tett semmilyen érdemi javaslatot. Ez még inkább megkérdőjelezte a puha klasszifikációs eredmények értékelhetőségét.

További probléma volt, hogy a másodlagos címkék kezelése sem volt egyértelmű. A liberális kappa kiszámításának módjából adódóan a legmagasabb értékhez az vezetett, ha a modell szinte minden esetben elfogadta másodlagos címkének a második legmagasabb valószínűséget kapott kategóriát. Ez viszont módszertanilag nem minden esetben indokolható, és a kappa értékek torzítottan magasabbnak tűnhettek a valós teljesítményhez képest. Ennek alternatívájaként lehetőség lett volna előre meghatározott küszöbérték alkalmazására (pl. csak akkor rendelünk másodlagos címkét, ha az értéke legalább 0,3 vagy 0,4). Azonban a puha klasszifikációs megközelítés esetében már a konzervatív kappa értékek is alacsonyabbnak bizonyultak, mint a kemény klasszifikáció során mért értékek, valamint a kemény megközelítés eredményei értelmezhetőbbek és egyértelműbbek is voltak. Ezen okokból kifolyólag a végső modellek kiértékeléséhez kizárólag a kemény klasszifikációs eredményeket használtam fel.

Végül tehát a kemény klasszifikációs megközelítésen belül a részletesebb zero-shot és a few-shot promptolást hasonlítottam össze az egyes modelleken belül, valamint a két modell között az egymásnak megfeleltethető elrendezéseket (2. táblázat).

2. táblázat A végső modell elrendezések tulajdonságai

Modell	Klasszifikációs megközelítés	Promptolási technika	Temperature értékek
GPT-4o mini	Kemény	Zero-shot (kategóriák részletes leírása)	0
		Few-shot (kategóriák + példák)	0
Llama 3.3 70B	Kemény	Zero-shot (kategóriák részletes leírása)	0
		Few-shot (kategóriák + példák)	0

5. Elemzés

A modellek klasszifikációs teljesítményének értékelésében egyaránt kitérek a pontosságukra és megbízhatóságukra. A pontosság esetében megvizsgálom a globális eredményeket, liberális és konzervatív módon, az eredeti öt kategóriás címkézés és négy kategóriás átkódolás mellett egyaránt. A kategóriánkénti pontosságot hasonlóan elemzem

majd, de részletesen csak az eredeti öt kategóriában. Ezen felül kitérek arra is, hogy az egyes modellek pontossága hogyan alakult a könnyű és nehéz esetek csoportjaiban. Végül azt is megvizsgálom, hogy a konfidencia pontszámok alapján kialakított három csoportban hogyan néz ki a modellek teljesítménye.

A pontosság után rátérek a modellek megbízhatóságának vizsgálatára. Egyrészt elemzem, hogy egyazon modell mennyire következetes különböző promptolási stratégiák esetén. Másrészt összevetem a GPT-4o mini és a LLaMA 3.3 válaszait az egymásnak megfeleltethető promptok mentén. Végül azt is bemutatom, hogy a modellek mennyire konzisztens módon működnek többszöri futtatás során.

5.1. Pontosság

5.1.1. Globális pontosság

Ebben a fejezetben a nagy nyelvi modellek globális pontosságának eredményeit mutatom be különböző klasszifikációs és promptolási technikák szerint, konzervatív és liberális értékelési szemléletmóddal, a precision, recall, F1-score és a Cohen-féle kappa segítségével. A modellektől a promptokban öt kategóriás címkézést kértem, a pontossági mutatókat azonban négy kategóriára is kiszámoltam, összevonva az irreleváns és nem besorolható címkéket. Ennek oka elsősorban az volt, hogy ha a modell csupán ezt a két címkét cserélte fel, az egy enyhébb hibának minősül, mintha az érvényes kategóriák annotálásában ront, így érdemes megvizsgálni, hogyan változik a kép, ha csak négy kategóriában vizsgáljuk az eredményeket.

Ha az eredeti öt kategóriát figyelembe véve vizsgáljuk a modellek eredményeit, akkor azt láthatjuk, hogy a zero-shot esetekben a konzervatív és liberális módon számolt Cohen-féle kappa és az F1-score egyaránt alacsonyabb, mint a few-shot megközelítés használata során, ezek az eltérések ugyanakkor nem kifejezetten jelentősek. A modelleknek nyújtott példák tehát jelen esetben növelték a pontosságot, ugyanakkor nem számottevő mértékben. A négy különböző esetben mind a liberális, mind a konzervatív kappa 0,4-0,6 között mozog, amely a bevett értelmezések szerint csak mérsékelt egyetértést jelent (Landis & Koch, 1977).

3. táblázat Globális pontossági mutatók öt kategóriára nézve konzervatív és liberális számítás mellett

	GPT zero-shot		GPT few-shot		Llama zero-shot		Llama few-shot	
	Konz.	Lib.	Konz.	Lib.	Konz.	Lib.	Konz.	Lib.
Cohen-féle kappa	0,42	0,52	0,45	0,54	0,45	0,57	0,48	0,58
Precision	0,65	0,71	0,65	0,70	0,64	0,70	0,61	0,68
Recall	0,57	0,64	0,58	0,66	0,61	0,70	0,63	0,70
F1-score	0,59	0,66	0,60	0,67	0,58	0,66	0,60	0,68

Szintén érdemes még kiemelni, hogy az egymásnak megfeleltethető promptokat figyelembe véve a Llama 3.3 a kappa tekintetében minimálisabban jobban teljesít, mint a GPT-4o mini, ez ugyanakkor nem nagy különbség és az F1-score esetében még minimálisabban áll fenn ez a tendencia, a két modell tehát szinte azonos teljesítményt nyújt. A liberális módon való számítás a kappában majdnem minden modell esetében 0,1-es javulást eredményez. Ez különösen azért érdekes, mert a GPT-4o mini jelentősen kevesebb érvényes másodlagos címkét (zero-shot: 998, few-shot: 853) ad, mint a Llama 3.3 (zero-shot: 2502, few-shot: 1482), ennek ellenére mégis hasonló javulást mutat, ez arra enged következtetni, hogy a Llama 3.3 sok esetben olyan másodlagos címkét ad, amely nem egyezik meg a valódi elsődleges címkével, más különben nagyobb javulást várhatnánk.

4. táblázat Globális pontossági mutatók négy kategóriára nézve konzervatív és liberális számítás mellett

	GPT zero-shot		GPT few-shot		Llama zero-shot		Llama few-shot	
	Konz.	Lib.	Konz.	Lib.	Konz.	Lib.	Konz.	Lib.
Cohen-féle kappa	0,48	0,58	0,49	0,59	0,46	0,58	0,49	0,59
Precision	0,66	0,71	0,66	0,72	0,68	0,74	0,66	0,72
Recall	0,63	0,70	0,64	0,71	0,62	0,71	0,64	0,71
F1-score	0,63	0,70	0,64	0,71	0,60	0,68	0,62	0,70

Négy kategóriába átkódolva az látszik, hogy a különböző megközelítések kappa értéke a GPT esetében minimálisan növekedett, míg a Llama esetében ez a növekedés nagyjából elhanyagolható mértékű. Ez alapján feltételezhetjük, hogy a modellek teljesítményét elsősorban nem az irreleváns és nem besorolható kategóriák felcserélése rontja el. A GPT és a Llama közötti különbségek azonban a négy kategóriás megközelítés mellett tovább csökkentek, sőt zero-shot promptolás mellett a GPT-4o mini jobban is teljesített, mint a Llama 3.3-as modellen futtatott párja.

Összeségében az eredményekből az látszódik, hogy mind a GPT-4o mini, mind a Llama 3.3 modellek esetében a few-shot promptolási megközelítés alkalmazása enyhe javulást eredményezett a globális pontosság mutatóiban, ez a javulás azonban nem volt igazán nagy mértékű. Emellett a konzervatív és liberális szemléletű értékelések közötti különbség arra utal, hogy bár a modellek sok esetben képesek helyes címkéket adni, de nem mindig sikerül helyesen döntést hozniuk az elsődleges és másodlagos címkék rangsorolásában. Az eredmények négy kategóriába történő átkódolása szintén együtt járt a pontosság növekedésével, ez azonban a Llama modellek esetében nagyon minimális mértékű volt, itt tehát feltehetőleg a hibák jelentős része nem az irreleváns és a nem besorolható kategóriák felcseréléséből adódott. Mindent egybevetve a globális pontossági mutatókat nézve a Llama 3.3 minimálisan jobban teljesített, mint a GPT-4o mini, a különbség azonban nem jelentős.

5.1.2. Kategóriák szerinti pontosság

Annak érdekében, hogy tisztább képet kapjak a modellek pontosságáról és döntéseikről, a teljesítményüket nem csak globálisan, hanem kategóriánként is megvizsgáltam. Ennek során külön figyelmet fordítottam arra, hogy mely kategóriák jelentenek nagyobb kihívást a modellek számára, és hogy milyen típusú hibák dominálnak a címkézések során.

5. táblázat Kategóriánkénti pontossági mutatók konzervatív és liberális számítás mellett (GPT zero-shot)

	Biomedikális		Pszichológiai		Szociológiai		Nem besorolható		Irreleváns	
	Konz.	Lib.	Konz.	Lib.	Konz.	Lib.	Konz.	Lib.	Konz.	Lib.
Cohen-féle kappa	0,62	0,72	0,42	0,57	0,34	0,55	0,05	0,06	0,44	0,44
Precision	0,85	0,88	0,56	0,69	0,39	0,53	0,08	0,08	0,73	0,73
Recall	0,61	0,71	0,73	0,78	0,42	0,67	0,21	0,23	0,44	0,44
F1-score	0,71	0,78	0,64	0,73	0,40	0,60	0,11	0,11	0,55	0,55

A GPT zero-shot megközelítés a legjobb eredményt a biomedikális szövegek esetében érte el, mind konzervatív, mind liberális kiértékelés mellett. Itt a Cohen-féle kappa értéke 0,62, illetve 0,72, ami azt jelzi, hogy a modell elég pontosan képes felismerni ezt a kategóriát. A precision értékek különösen magasak, vagyis a modell által biomedikálisnak besorolt posztok túlnyomó többsége valóban ide tartozik. Ugyanakkor a recall értékek ennél alacsonyabbak, ami

arra utal, hogy bár a modell pontosan címkéz, viszont nem minden ide tartozó szöveget talál meg.

A pszichológiai szövegek esetén más jellegű mintázat figyelhető meg, itt a recall értékek magasabbak, vagyis a modell az esetek nagy részében helyesen felismeri az ilyen tartalmakat. A precision ugyanakkor valamivel alacsonyabb, ami arra utal, hogy a pszichológiai címkét olyankor is gyakran használja, mikor nem kellene. A kappa értékek (0,42; 0,57) ezzel összhangban közepes eredményt mutatnak.

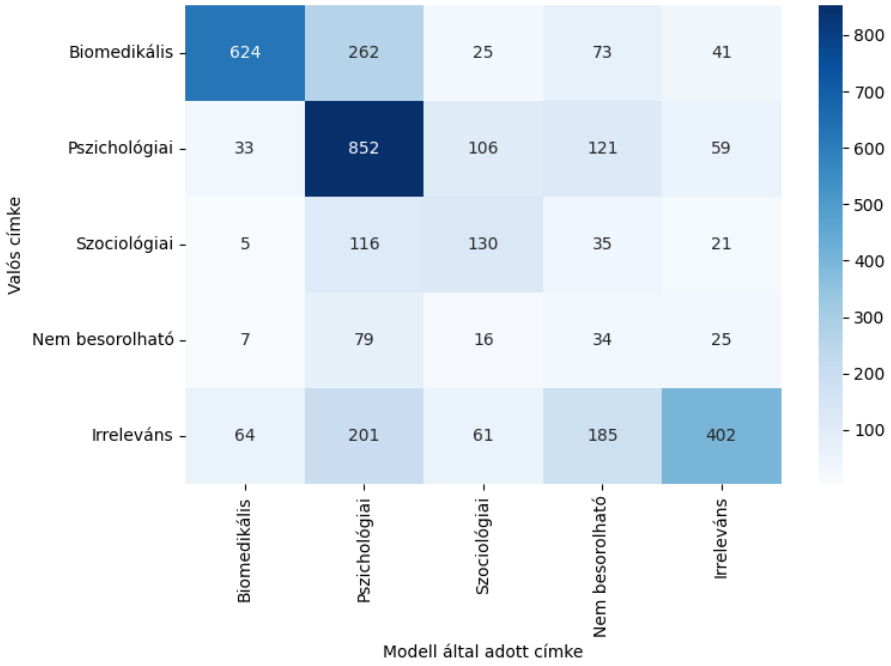
A szociológiai posztok felismerése jelentette a legnagyobb kihívást a három keretezési kategória közül. A konzervatív kappa érték 0,34, a liberális pedig 0,55, előbbi egy elég alacsony értéknek mondható, míg utóbbi inkább közepes. A recall értékek alapján a modell gyakran nem ismeri fel ezeket a posztokat, és a precision-ből kiindulva azok közül is, amiket ide sorol, sok valójában más kategóriába tartozik. Mindez arra utal, hogy a keretezések közül a szociológiai kategória besorolása bizonyult a legnehezebbnek a modell számára.

A nem besorolható kategória esetében mind a precision, mind a recall értékek kifejezetten alacsonyak, és a Cohen-féle kappa értéke is nagyon gyenge. Ez arra utal, hogy a modell csak ritkán ismeri fel helyesen az ide tartozó posztokat, és amikor mégis ezt a címkét alkalmazza, az is sokszor téves. Az irreleváns kategória ezzel szemben valamivel jobban teljesít. A recall értékek itt is mérsékeltek, vagyis a modell nem mindig ismeri fel, ha egy poszt teljesen irreleváns. Ugyanakkor a precision meglepően magas, vagyis ha mégis ezt a címkét adja egy posztra, az jellemzően valóban ebbe a kategóriába tartozik.

A modellek címkézési teljesítményét confusion mátrixok segítségével is megvizsgáltam, ebben az esetben a könnyebb érthetőség érdekében a konzervatív megközelítéssel dolgoztam, amely az elsődleges címkéket hasonlítja csak össze. Az ábra alapján a biomedikális címkénél nyújtott teljesítmény valóban kiemelkedő a többihez képest. A modell viszont kifejezetten gyakran címkézi pszichológiainak az ide tartozó szövegeket. Szintén jól látszik, hogy a valójában pszichológiai keretezés említő posztokat hajlamos a modell szociológiainak vagy nem besorolhatónak azonosítani. Hasonlóan a szociológiai kategóriát pedig leggyakrabban a pszichológiaival keveri össze. Mindez jól mutatja, hogy sok esetben a modellnek is nehéz elkülöníteni az egyes kategóriákat, és hogy az egyes keretezéseknek lehetnek hasonló vagy átfedő részei, amelyek nehezítik a döntéshozatalt. Emellett pedig az is jól látszódik, hogy

irreleváns címke helyett a modell gyakran ad nem besorolhatót, ez viszont fordítva sokkal ritkább esetben áll fenn.

2. Ábra A valós elsődleges és a modell által adott elsődleges címkék confusion mátrixa (GPT zero-shot)



A GPT few-shot technikát alkalmazó verziója szinte ugyanezeket a tendenciákat mutatja. A különbség csak annyi, a szociológiai kategóriát leszámítva minden egyéb kategóriában minimálisan nagyobb pontosságot sikerült elérnie. Az ide tartozó confusion mátrix szintén nagyon nagy hasonlóságot mutat a GPT zero-shot eredményéhez (3. számú függelék).

6. táblázat Kategóriánkénti pontossági mutatók konzervatív és liberális számítás mellett (GPT few-shot)

	Biomedikális		Pszichológiai		Szociológiai		Nem besorolható		Irreleváns	
	Konz.	Lib.	Konz.	Lib.	Konz.	Lib.	Konz.	Lib.	Konz.	Lib.
Cohen-féle kappa	0,64	0,74	0,45	0,59	0,31	0,52	0,07	0,08	0,45	0,45
Precision	0,84	0,86	0,59	0,71	0,34	0,52	0,09	0,09	0,70	0,70
Recall	0,64	0,75	0,71	0,78	0,42	0,63	0,22	0,24	0,48	0,48
F1-score	0,73	0,80	0,65	0,74	0,38	0,57	0,13	0,13	0,57	0,57

A Llama zero-shot modell esetében több ponton hasonlóságot mutat a GPT zero-shot eredményeivel, ugyanakkor néhány eltérés is megfigyelhető. A biomedikális és pszichológiai kategóriák itt is jól teljesítenek, különösen a liberális számítás mellett. A recall értékek ezekben a kategóriákban kiemelkedőek, például a pszichológiai címkéknél mindkét értékelési módszerrel 0,86–0,87 körüli értéket látunk, ami magasabb, mint a GPT esetében, a biomedikális kategória esetében pedig a liberális érték (0,91) kifejezetten magas. Bár ezzel együtt a precision pedig kicsivel rosszabb. A szociológia címke eredményei gyengébbnek bizonyultak a Llama esetében, különösen konzervatív megközelítés mellett (kappa = 0,29).

7. táblázat Kategóriánkénti pontossági mutatók konzervatív és liberális számítás mellett (Llama zero-shot)

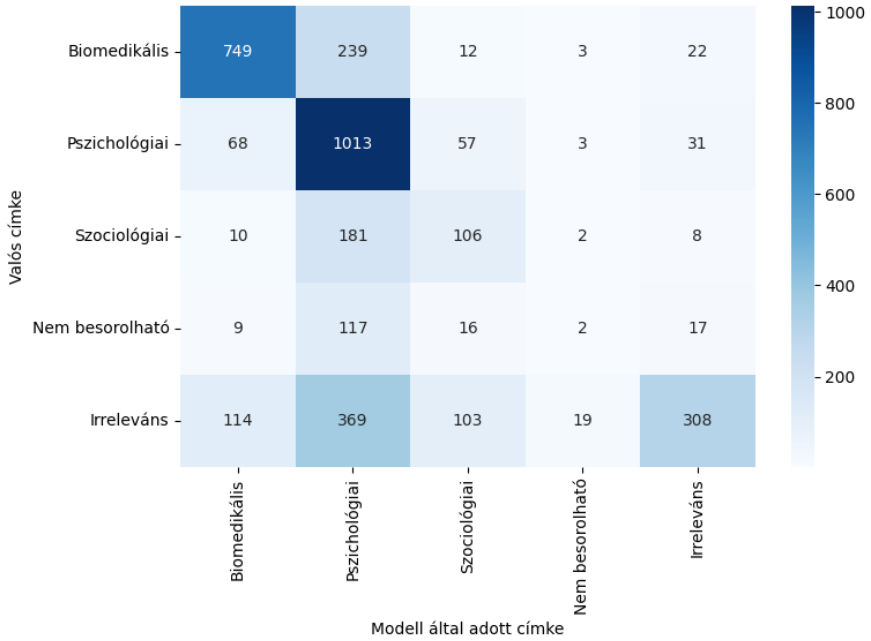
	Biomedikális		**Pszichológiai**		**Szociológiai**		**Nem besorolható**		**Irreleváns**	
	Konz.	**Lib.**	**Konz.**	**Lib.**	**Konz.**	**Lib.**	**Konz.**	**Lib.**	**Konz.**	**Lib.**
Cohen-féle kappa	0,67	0,79	0,42	0,57	0,29	0,52	0,01	0,01	0,38	0,38
Precision	0,79	0,80	0,53	0,66	0,36	0,48	0,07	0,07	0,80	0,80
Recall	0,73	0,91	0,86	0,87	0,35	0,65	0,01	0,01	0,34	0,34
F1-score	0,76	0,85	0,66	0,75	0,35	0,55	0,02	0,02	0,47	0,47

A legpontatlanabb eredmény továbbra is a nem besorolható kategóriában mutatkozik meg, itt a kappa, precision, recall és F1-score értékek egyaránt nagyon alacsonyak, ennek egyik oka az lehet, hogy a modell összesen csak 29 esetben használja ezt a címkét és az eredmények alapján akkor is rosszul. Egy fokkal pontosabb az irreleváns kategória, itt GPT-hez hasonlóan a precision kifejezetten jó, tehát ha a modell ezt a címkét adja, az rendszerint pontos, azonban a recall alapján sok esetben nem használja a kategóriát mikor egyébként kellene.

A Llama zero-shot promptolása esetében a confusion mátrix sok tekintetben hasonlít a GPT-nél látott eredményekhez. A biomedikális kategória helyett ez a modell is sokszor prediktál pszichológiai címkét, illetve a szociológiai és pszichológiai címkék mindkét irányú összekeverése szintén egy olyan hiba, amely továbbra is fennáll, bár a valós pszichológiai keretezésű szövegeket ez a modell kevésbé sorolja be a szociológiai kategóriába. A nem besorolható kategória rossz teljesítményében pedig nagy szerepet játszik, hogy a modell hajlamos ezeket a posztokat pszichológiai keretezésűként azonosítani. Ez a hiba már a GPT esetében is feltűnő, de itt még inkább jellemző. Ez a tendencia pedig az irreleváns kategória

esetében szintén szembetűnő. Jól látszik, hogy ezeknél a kategóriáknál az elsődleges gond ebben az esetben nem az egymástól való megkülönböztetésükből fakad, hanem a többi kategóriára adott hamis pozitív eredményeikből.

3. Ábra A valós elsődleges és a modell által adott elsődleges címkék confusion mátrixa (Llama zero-shot)



A Llama few-shot modell teljesítménye hasonló tendenciát mutat, mint a zero-shot esetben. A biomedikális és a pszichológiai kategóriákban továbbra is jól teljesít. A biomedikális szövegek esetében a konzervatív precision romlott, ezzel együtt a recall viszont javult. A pszichológiai keretezésnél pedig ennek fordítottja áll fenn.

A szociológiai kategóriánál továbbra is kifejezetten rosszul teljesít, bár ebben az esetben jellemzőbb az, hogy ha ezt a kategóriát adja a modell az egyezik a valós címkével, a recall viszont romlott, tehát nem ismeri fel elég jól a valóban szociológiai keretezést tartalmazó posztokat.

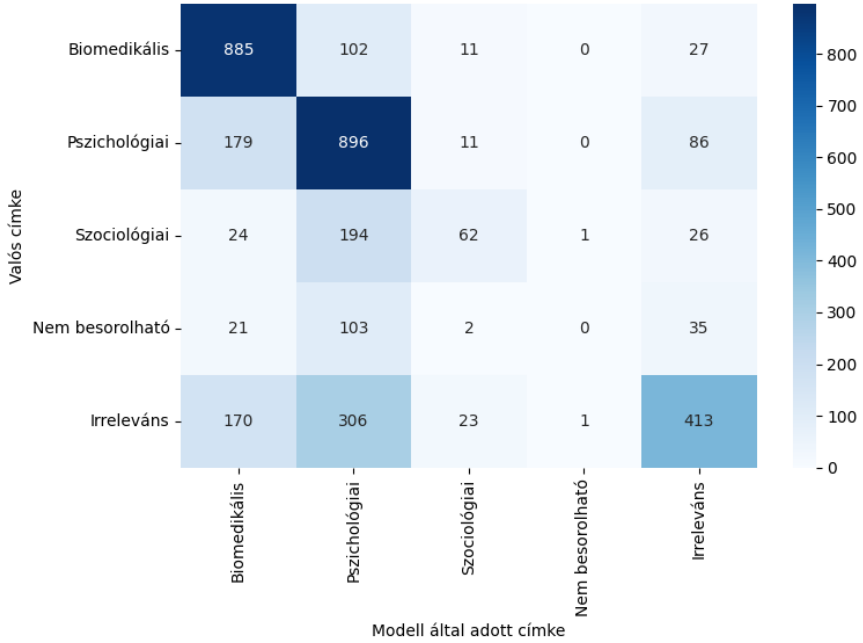
A legnagyobb problémát továbbra is a nem besorolható kategória jelenti, ahol gyakorlatilag nulla értékeket látunk minden metrikában. Az irreleváns kategóriában ezzel szemben a kappa értéke nőtt, 0,38-ról, 0,44-re, viszont a precision csökkent, a recall azonban szintén nőtt. Tehát többet ismert fel helyesen a modell ebből a kategóriából, de többször is sorolta helytelenül ide a szövegeket.

8. táblázat Kategóriánkénti pontossági mutatók konzervatív és liberális számítás mellett (Llama few-shot)

	Biomedikális		Pszichológiai		Szociológiai		Nem besorolható		Irreleváns	
	Konz.	Lib.	Konz.	Lib.	Konz.	Lib.	Konz.	Lib.	Konz.	Lib.
Cohen-féle kappa	0,66	0,75	0,43	0,58	0,27	0,52	0,00	0,00	0,44	0,44
Precision	0,69	0,78	0,56	0,66	0,57	0,59	0,00	0,00	0,70	0,70
Recall	0,86	0,89	0,77	0,85	0,20	0,51	0,00	0,00	0,45	0,45
F1-score	0,77	0,83	0,65	0,74	0,30	0,55	0,00	0,00	0,55	0,55

A Llama esetében a few-shot promptolási megközelítés a confusion mátrix tekintetében kissé más eredményekre vezetett, bár az általános tendenciák megmaradtak.

4. ábra A valós elsődleges és a modell által adott elsődleges címkék confusion mátrixa (Llama few-shot)



A biomedikális kategória esetében, azonban a modell kevesebbszer adott hamisan pszichológiai címkét a szövegeknek. Hasonlóan, a pszichológiai keretezésre kevesebbszer adott szociológiai címkét, viszont a korábbinál sokkal többször választotta a biomedikális címkét hamisan. A globális mutatók alapján korábban már láttuk, hogy a few-shot technika kicsit javított a szövegek besorolásán, de a confusion mátrixból az is kiolvasható, hogy ez az átfogó javulás bizonyos kategóriák és aspektusok esetében az eredmények romlásával járt együtt.

A kategóriánkénti részletes elemzésből tehát kiderült, hogy a modellek jelentős különbségekkel teljesítenek az egyes keretezési típusokban. A biomedikális kategóriában a többihez képes kiemelkedően teljesítenek a modellek és a pszichológiai címke esetében is jó eredményeket produkáltak, ezzel szemben a szociológiai, illetve különösen a nem besorolható kategória már sokkal nagyobb kihívást jelentett. Az irreleváns kategória felismerése pedig szintén vegyes eredményeket hozott.

5.1.3. Pontosság a könnyű és nehéz eseteknél

Az előző fejezetekben részletesen bemutattam a modellek globális és kategóriánkénti teljesítményét, azonban érdemes azt is megvizsgálni, hogy ezek a modellek hogyan teljesítenek olyan esetekben, amelyek az emberi annotátorok számára könnyen vagy éppen nehezen besorolhatók voltak. Ebben a fejezetben ezért külön összevetem a modellek pontosságát a két eltérő nehézségű szövegcsoportban. Az elemzéshez a már korábban alkalmazott konzervatív és liberális pontossági metrikákat alkalmazom.

9. táblázat Globális pontossági mutatók öt kategóriára nézve konzervatív és liberális számítás mellett a könnyű és nehéz esetek csoportjaiban

		GPT zero-shot		GPT few-shot		Llama zero-shot		Llama few-shot	
		Konz.	Lib.	Konz.	Lib.	Konz.	Lib.	Konz.	Lib.
Könnyű esetek	**Cohen-féle kappa**	0,58	0,66	0,60	0,68	0,61	0,70	0,64	0,71
	Precision	0,75	0,80	0,75	0,80	0,74	0,78	0,71	0,77
	Recall	0,69	0,75	0,71	0,76	0,73	0,79	0,75	0,80
	F1-score	0,71	0,77	0,72	0,78	0,72	0,77	0,72	0,78
Nehéz esetek	**Cohen-féle kappa**	0,23	0,33	0,24	0,35	0,24	0,40	0,27	0,40
	Precision	0,50	0,58	0,51	0,59	0,49	0,58	0,48	0,56
	Recall	0,42	0,51	0,43	0,52	0,45	0,57	0,48	0,58
	F1-score	0,42	0,51	0,44	0,53	0,41	0,52	0,44	0,54

A táblázatból az látszik, hogy minden modell és minden promptolási eljárás esetében lényegesen jobb eredmények születtek a könnyen besorolható szövegeknél, mint a nehezen kategorizálhatóknál. Ez a különbség különösen a Cohen-féle kappa értékeknél szembetűnő, a könnyű eseteknél a kappa értékek jellemzően 0,58 és 0,71 közötti tartományban mozognak, míg a nehéz eseteknél ez jelentősen alacsonyabb, 0,23 és 0,40 közötti értékeket mutat.

A többi pontossági mutató is hasonló mintázatot követ. A könnyű esetek esetében a precision jellemzően 0,71 és 0,80 között változik, a recall pedig 0,69 és 0,80 között mozog, ami azt mutatja, hogy a modellek többsége megbízhatóan képes azonosítani az egyértelműbb kategóriákat. Ezzel szemben a nehéz eseteknél a precision értékek alacsonyabbak (0,41–0,59), és a recall értékek is mérsékeltebbek (0,42–0,58). Ez arra utal, hogy a modellek nehezebben azonosítják be az egyes kategóriákat és az adott címkéik is gyakran hamis pozitívok.

Az eredmények alapján azt mondhatjuk, hogy az emberi annotátorok számára könnyűnek bizonyuló szövegek a nagy nyelvi modellek számára is egyszerűbben bekategorizálhatóak voltak. Elképzelhető tehát, hogy a modellek és az emberek is hasonló mintázatok és nyelvi jellegzetességek alapján döntenek a korpuszban szereplő szövegek kategorizációja során.

5.1.4. Pontosság és konfidencia pontszám

Hasonlóan a könnyű és nehéz esetek vizsgálatához ebben a fejezetben arra térek ki, hogy a modellek által adott konfidencia pontszámokból létrehozott három csoportban vajon eltérően alakulnak-e a különböző pontossági metrikák. Ebben az esetben a kérdés lényege, hogy a modell vajon helyesen ismeri-e fel a saját teljesítményét és tényleg nagyobb pontossággal kategorizálja-e azokat az eseteket, amelyekhez magas pontszámokat rendel.

A konfidencia pontszámokat három kategóriába soroltam: alacsony (0-79), közepes (80-89) és magas (90-100). Az intervallum egyenlőtlen felosztását az indokolta, hogy a modellek jellemzően sokkal több magas pontszámot adtak, mint alacsonyat és a 80 alatti értékeket alig használták ki. Ez azt eredményezte, hogy a kategóriák között nagy eltérések jelentkeztek a mintaszámokat tekintve, például a GPT zero-shot modell viszonylag jelentős számú (876) alacsony konfidenciájú esetet sorolt fel, míg a Llama modellek, különösen a few-shot verzióban, alig adtak ilyen értékelést (mindössze 42 és 3 eset). Természetesen a Llama modelleknél mindez azzal jár, hogy az alacsony konfidenciájú csoportban kiszámolt eredmények megbízhatósága alacsony lehet. Ennek ellenére azért maradtam a hármas csoportbontásnál, mert egy 0-ás és 85-ös megbízhatósági pontszám összevonása értelmezhetőségi szempontból nehezen indokolható.

Az eredmények alapján viszont így is egy egyértelmű tendencia rajzolódik ki, minél magasabb volt a modell által jelzett konfidencia, annál jobbak voltak a pontossági mutatók. Ez a mintázat különösen jól látszik a Cohen-féle kappa értékekből, amelyek az alacsony

konfidenciájú csoportban nagyon alacsony értékeket mutattak. A közepes konfidenciájú csoportokban azonban már javulás volt megfigyelhető, ami a magas konfidenciájú csoportokban tovább emelkedett, különösen a Llama modellek esetében, ahol már a konzervatív kappa is meghaladta a 0,6-os határt. Szintén érdekes, hogy a konzervatív és liberális kappák közötti különbség a közepes pontszámmal rendelkező csoportban a legnagyobb, ami utalhat arra is, hogy a modellek bizonytalanságának egy része abból fakadt, hogy az általuk megadott kategóriák sorrendjében nem voltak biztosak.

10. táblázat Globális pontossági mutatók öt kategóriára nézve konzervatív és liberális számítás mellett a különböző konfidencia pontszámú csoportokban

		GPT zero-shot		GPT few-shot		Llama zero-shot		Llama few-shot	
		Konz.	Lib.	Konz.	Lib.	Konz.	Lib.	Konz.	Lib.
Alacsony	**Cohen-féle kappa**	0,15	0,19	0,17	0,18	-0,01	0,06	0,00	0,00
	Precision	0,49	0,51	0,60	0,62	0,03	0,11	0,11	0,11
	Recall	0,27	0,30	0,30	0,31	0,07	0,14	0,33	0,33
	F1-score	0,27	0,30	0,34	0,35	0,04	0,11	0,17	0,17
Közepes	**Cohen-féle kappa**	0,33	0,49	0,32	0,46	0,26	0,42	0,21	0,33
	Precision	0,52	0,59	0,55	0,61	0,35	0,42	0,31	0,38
	Recall	0,56	0,68	0,53	0,62	0,47	0,59	0,47	0,57
	F1-score	0,51	0,62	0,50	0,59	0,38	0,49	0,35	0,46
Magas	**Cohen-féle kappa**	0,56	0,67	0,57	0,67	0,64	0,72	0,60	0,68
	Precision	0,68	0,75	0,69	0,76	0,74	0,78	0,70	0,75
	Recall	0,68	0,76	0,69	0,76	0,75	0,80	0,72	0,78
	F1-score	0,67	0,75	0,68	0,76	0,74	0,79	0,70	0,76

5.2. Megbízhatóság

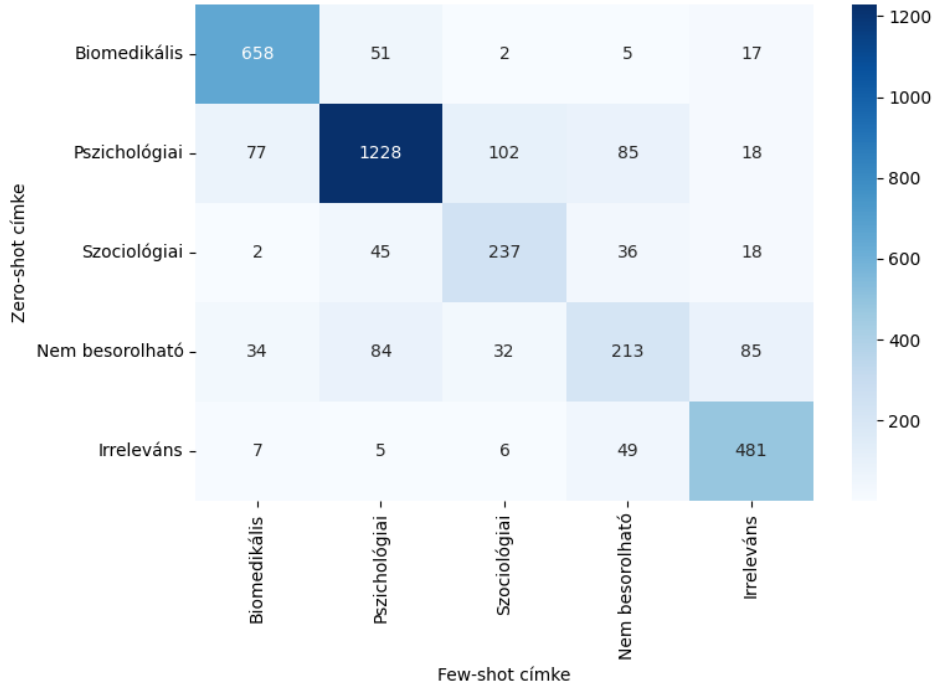
5.2.1. Azonos modell és különböző prompt

A pontosság vizsgálata mellett a modellek megbízhatóságára is kitérek, aminek elemzéséhez szintén a Cohen-féle kappát használtam fel. A kérdéskört pedig több szempontból is megközelítettem. A megbízhatóságot egyrészt az alapján megvizsgáltam, hogy ugyanaz a modell a két különböző promptolási stratégia mellett mekkora egyetértésre jut. Ez segíthet annak megválaszolásában, hogy a modellek mennyire érzékenyek a promptolásra.

A GPT-4o mini esetében a zero-shot és a few-shot megközelítés közötti konzervatív kappa értéke 0,71, ez egy viszonylag magas egyetértést tükröz. A liberális kappa pedig 0,76, a

növekedés arra enged következtetni, hogy vannak olyan esetek, ahol az elsődleges címkében eltérő volt a két megközelítés eredménye, de a másodlagos címkék bevonásával volt egyezés a két modell kategorizációjában. Ha megnézzük a konzervatív értelemben kirajzolható „confusion mátrixot", akkor láthatjuk, hogy valóban nagy az egyetértés a két megközelítés eredményei között. Ebben az esetben is a biomedikális és a pszichológiai kategória tűnik a legegyértelműbbnek, míg például a szociológiai keretezés tekintetében azért van egyet nem értés a két megközelítés között, ahogy a nem besorolható és irreleváns kategóriák esetében is.

5. ábra Az elsődleges címkék összevetése a GPT zero-shot és few-shot verziói között

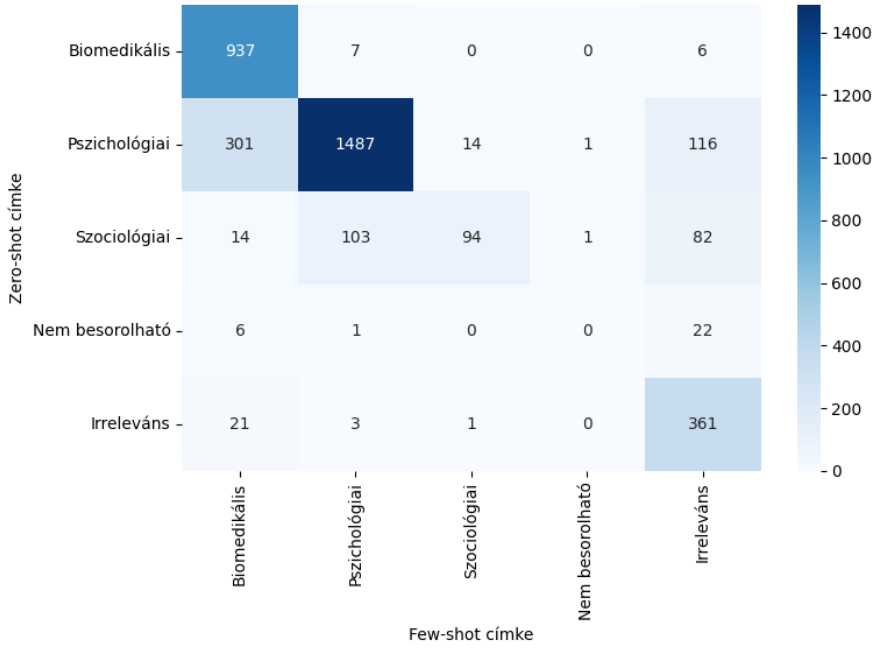


Érdemes lehet mindezt négy kategóriába is összevonni, hiszen ha a két megközelítés például az irreleváns és nem besorolható kategóriák megkülönböztetésében nem ért egyet, az nem feltétlenül nagy probléma. Ebben az esetben a konzervatív kappa értéke 0,75, míg a liberálisé 0,8, az egyetértés tehát növekedett az eredeti öt kategóriás esethez képest, a javulás azonban csak 0,04-es, vagyis nem ennek a két kategóriának a felcserélése az, ami lényeges rontja az eredményt.

A Llama 3.3 esetében ugyanezt vizsgálva, öt kategóriában a konzervatív kappa 0,7, a liberális pedig 0,72. Ezek a GPT-hez képest alacsonyabb érték, ugyanakkor nem jelentős az eltérés. Érdekes még, hogy a Llama esetében a konzervatív és liberális értékek között különbség alacsonyabb, így ebben az esetben nem segített annyi a másodlagos kategóriák

figyelembevétele az egyetértés növelésében, mint a GPT-nél. A „confusion mátrixot” megnézve azt láthatjuk viszont, hogy a modellek közötti egyetértés elég nagy, sőt a mátrix alapján elsőre jobbnak is tűnhet, mint a GPT esetében. Azonban viszonylag sok olyan eset van, amelyet a zero-shot megközelítés pszichológiainak címkézett, míg a few-shot biomedikálisnak, hasonló eset áll fenn a zero-shot modell szerint szociológiai keretezést tartalmazó szövegek esetében is, itt a few-shot gyakran inkább pszichológiainak gondolta őket. Ezen felül pedig a few-shot megközelítés több irreleváns címkét is adott, mint a zero-shot. A kappa értékét pedig tovább ronthatja, hogy a nem besorolható kategóriát nézve egyetlen szöveg esetében sem értett egyet a két promptolási módszer.

6. Ábra Az elsődleges címkék összevetése a Llama zero-shot és few-shot verziói között



Négy kategóriába átkódolva az eredményeket a Llama esetében még kisebb javulás tapasztalható, mint a GPT-nél. A konzervatív kappa értéke 0,71, a liberálisé pedig 0,73. Ez ugyanakkor nem meglepő, hiszen már a mátrixból is jól látszódott, hogy az egyet nem értés elsődleges oka nem az irreleváns és nem besorolható kategóriák felcserélése.

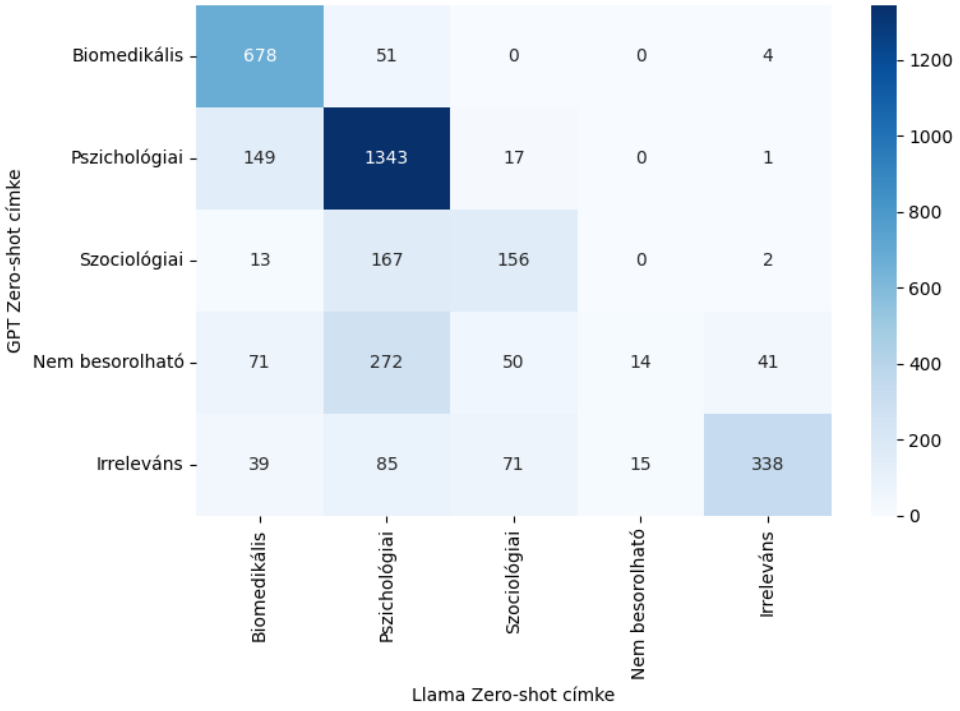
5.2.2. Azonos prompt és különböző modell

Csak úgy, mint az emberi annotátorok esetében, itt is fontos kérdés, hogy a különböző modellek ugyanazon prompt esetében mekkora egyetértésre jutnak. Ez a felállás tulajdonképpen hasonlóan működik, mint a hagyományos annotáció esetében, vajon, ha két

modell ugyanazokat az utasítások kapja, ahogy két annotátor az útmutatóban, akkor a döntéseik is hasonlóak lesznek?

Ha a GPT-4o mini és a Llama 3.3 zero-shot promptolását hasonlítjuk össze, akkor a konzervatív kappa értéke 0,58, míg a liberálisé 0,67. Ezek az értékek valamivel alacsonyabbak, mint mikor ugyanazon modellek eltérő promptolását vizsgáltuk meg, azonban konzervatív esetben már majdnem magas egyetértésről beszélhetünk, a liberális megközelítést nézve pedig el is érték ezt a modellek. Ha megnézzük a „confusion mátrixok", akkor azt látjuk, hogy a Llama szinte soha nem címkézet irrelevánsnak vagy nem besorolhatónak olyan szöveget, amelyet a GPT érvényes kategóriába sorolt, ugyanez fordítva viszont egészen gyakran megtörtént, főleg a Llama által pszichológiai keretezésűként azonosított szövegek esetében. Ugyanitt a GPT gyakran inkább szociológiainak gondolta az egyes posztokat.

7. Ábra Az elsődleges címkék összevetése a GPT és a Llama zero-shot verziói között

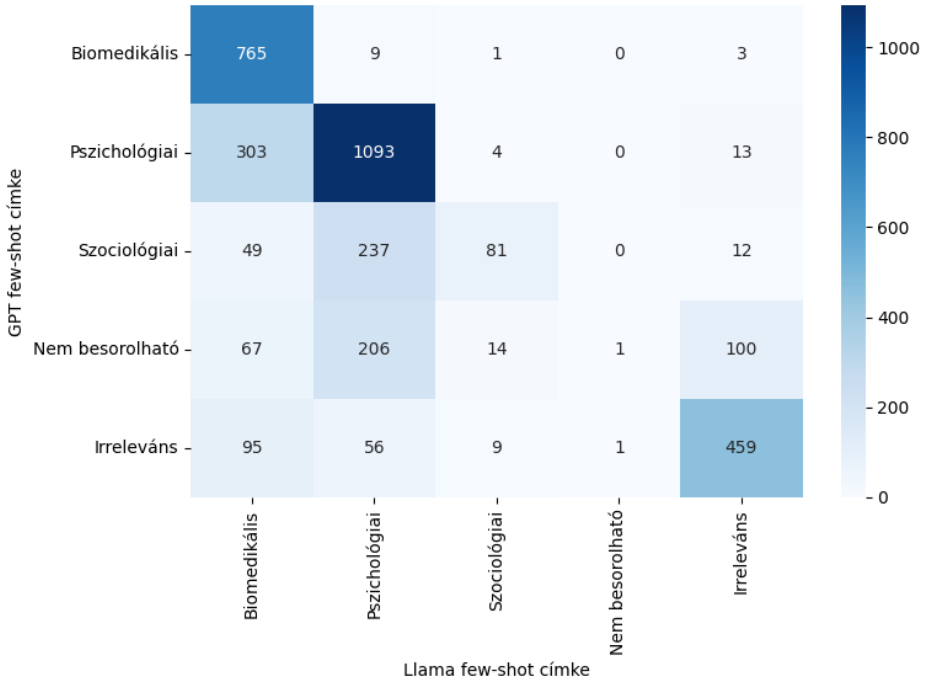


Négy kategóriába átsorolva a konzervatív kappa 0,59, míg a liberális 0,68. A javulás viszonylag alacsony mértéke itt sem meglepő, hiszen jól látszik, hogy az egyet nem értés itt sem az irreleváns és nem besorolható kategóriák felcserélésében nyilvánul meg.

A few-shot promptolás esetében az eredeti öt kategóriára vizsgálva a konzervatív kappa értéke 0,54, míg a liberálisé 0,66. Ezek valamivel alacsonyabb értékek, mint a zero-shot esetben, főleg a konzervatív kappát nézve. A „confusion mátrixot" megnézve hasonló kép

rajzolódik ki, mint a zero-shot promptolás mellett. A Llama modell ebben az esetben sem társít irreleváns vagy nem besorolható jelzőt, azokhoz a címkékhez, amelyek a GPT szerint a három keretezés valamelyikébe tartoznak, míg fordítva ez nem áll fenn. Emellett a Llama a GPT által pszichológiainak címkézett szövegek jelentős részét tartotta biomedikálisnak, míg a szociológiainak vagy nem besorolhatónak vélt posztoknak gyakran adott pszichológiai címkét.

8. Ábra Az elsődleges címkék összevetése a GPT és a Llama few-shot verziói között



Négy kategóriába átsorolva a konzervatív kappa értéke a few-shot eseteket vizsgálva 0,57, míg a liberálisé 0,69. Ebben az esetben a zero-shot promptoláshoz képest ez a növekedés kicsivel nagyobb, ami a confusion mátrix alapján várható is volt, hiszen látszik, hogy van összesen 101 olyan eset, amelyben az irreleváns és nem besorolható kategóriák összevonásával egyetértés keletkezik a modellek között.

Összeségében az látszik, hogy a különböző modellek azonos promptolásának eredményeit összehasonlítva az egyetértés alacsonyabb, mint ugyanazon modellek különböző promptjainak esetében. Az egyetértés konzervatív értelemben közepes, liberális megközelítéssel vizsgálva azonban már magasnak mondható mind a zero-shot, mind a few-shot esetben.

5.2.3. Azonos modell és prompt

Végül a modellek megbízhatóságát és konzisztenciáját az alapján is megvizsgáltam, hogy többször újrafuttatva ugyanazt a promptot mennyire jutnak ugyanarra az eredményre. Ehhez

összesen háromszor futtattam le minden elrendezést a teljes teszthalmazon, majd ezekre páronként kiszámoltam a Cohen-féle kappa két verzióját, majd vettem ezek átlagát. Itt fontos ismét megemlíteni, hogy a végleges modelljeim mindegyike nullás temperature beállítást kapott, amely elméletben a lehető legdeterminisztikusabb eredményekre kell, hogy vezessen. Ha megnézzük a modellek teljesítményét, akkor az látszódik, hogy a kappa mindkét verziója kiemelkedően magas értékeket mutat. A GPT zero-shot verziójában mutatkozott a legtöbb eltérő eset, ezt követte a GPT few-shot verziója. A két Llama elrendezés esetében mindkét érték 0,99.

11. táblázat A páronkénti konzervatív és liberális kappák átlaga globálisan, öt kategóriára nézve

	GPT zero-shot	GPT few-shot	Llama zero-shot	Llama few-shot
Konzervatív kappa	0,93	0,95	0,99	0,99
Liberális kappa	0,94	0,96	0,99	0,99

Az eredmények alapján azt mondhatjuk el, hogy habár a modellek nem viselkednek teljesen determinisztikusan (ami a nagy nyelvi modellek működéséből adódik még a nullás temperature érték mellett is), azonban így is kiemelkedő konzisztenciát és megbízhatóságot mutatnak abban az esetben, ha ugyanazokon az adatokon, ugyanazzal a prompttal és beállítással futtatjuk őket.

6. A kutatás korlátai

Természetesen a szakdolgozatomban elvégzett kutatásnak is megvoltak a maga korlátai. Ezek közül az egyik már a modellek kiválasztása esetén is fennállt. A számos különböző nagy nyelvi modell közül mind költségbeli, mint pedig időbeli korlátokból fakadóan csak a Llama 3.3 70B és a GPT-4o mini modellek teljesítményét vizsgáltam meg. Időbeli korlátok alatt ebben az esetben a modellek futtatásának időigényessége értendő. A teszthalmazban szereplő nagyjából 3700 szöveg estében például a GPT-4o mini egyetlen futtatása megközelítőleg egy órát vett igénybe. A kevés vizsgált modellből adódóan a nagy nyelvi modellek összeségére nem fogalmazhatók meg általánosítások az eredmények alapján. Emellett érdemes azt is kiemelni, hogy ezek nem a legnagyobb paraméterszámú, vagy legnagyobb tanító adattal rendelkező modellek, így könnyen lehetséges, hogy más LLM-ek pontosabb eredményre vezettek volna. Szintén az időbeli és költségbeli korlátokra vezethető vissza az is, hogy a különböző promptolási

és klasszifikációs megközelítések egy részét csak a tanítóhalmazon vizsgáltam meg és nem tudtam részletesebben elemezni a teljesítményüket a teszthalmazon. A további limitációk közé tartozik még, hogy a nagy nyelvi modellek annotációs teljesítményét csak egy adathalmazon vizsgáltam meg.

A dolgozatom egy másik korlátja részben a nagy nyelvi modellek működéséből fakad. Ezek a modellek gyakran feketedobozként működnek, így a döntéseik okairól keveset tudunk elmondani, ami az értelmezhetőséget csorbíthatja. Ebből fakadóan elsősorban én is csak a pontosságra és a megbízhatóságra fektettem a hangsúlyt. Ahogy azonban azt Németh (2024) is megjegyzi, az értelmezhetőség kérdésköre mindenképpen egy olyan téma, amelyre a jövőben nagyobb hangsúly fektetendő.

7. Összegzés

A szakdolgozatomban a nagy nyelvi modellek szöveganalitika felhasználhatóságával foglalkoztam. Ezen belül azt vizsgáltam, hogy egy depresszióval kapcsolatos annotált szövegkorpusz esetében mennyire képesek ezek a modellek pontosan replikálni az egyes posztok és bejegyzések címkéit. A szövegek ötféle kategóriába sorolhatók voltak: biomedikális, pszichológiai, szociológiai, irreleváns és nem besorolható. A posztok pedig elsődleges és szükség esetén másodlagos címkéket is kaphattak az emberi annotátoroktól.

A vizsgálat során a GPT-4o mini és a Llama 3.3 70B modellek teljesítményét hasonlítottam össze. Előbbi egy zárt forráskódú modell, míg utóbbi nyílt forráskódú. A felhasznált modelleket pedig API-n keresztül futtattam.

A kutatás során a tanítóhalmazon négy különböző temperature beállítást (0; 0,2; 0,4; 0,6), két klasszifikációs megközelítést (puha és kemény) és három promptolási technikát (rövid zero-shot, részletes zero-shot és few-shot) alkalmaztam. Az előzetes eredmények arra mutattak, hogy jelen eseteben a temperature beállítás alacsonyan tartása nem ment a pontosság rovására, így a végső modelleket a megbízhatóság növelése érdekében nullás érték mellett futtattam. A két klasszifikációs megközelítésből végül csak a kemény verziót alkalmaztam, míg a promptok közül a részletes zero-shot és few-shot technikákat.

A modellek teljesítményének értékelésekor a pontosságra és a megbízhatóságra egyaránt kitértem. A pontosság esetében olyan metrikákat alkalmaztam, mint a precision, recall, F1-

score, illetve a Cohen-féle kappa, amelyeket konzervatív és liberális módon egyaránt kiszámítottam. A promptolási technikák közül a few-shot mind a GPT, mind a Llama esetében jobb eredményre vezetett, viszont nem lényegesen magasabbra. Hasonlóan, a globális eredményeket nézve a Llama 3.3 70B jobban teljesített, mint a GPT-4o mini, a különbség viszont itt szintén alacsony. A globális kappák konzervatív esetben 0,42 és 0,48, míg liberális esetben 0,52 és 0,58 között mozogtak, amelyek inkább csak közepes pontosságnak számítanak. A kapott eredmények konzervatív kappa értékei nagyjából ugyanabban a tartományban mozognak, mint a Németh és munkatársai (2022) által alkalmazott hagyományos gépi tanulási módszereké. A nagy nyelvi modellek alkalmazása tehát ilyen szempontból előrelépést nem hozott.

A modellek pontosságát kategóriánként is kiértékeltem. Ebben az esetben szintén hasonló eredményre jutottam, mint Németh, Sik és Máté (2020), illetve Németh és munkatársai (2022). Rendszerint a nagy nyelvi modellek számára is a biomedikális kategória bizonyult a legkönnyebben besorolhatónak, amelyet pedig a pszichológiai követett. A szociológiai keretezés esetében a modellek teljesítménye már sokkal inkább megkérdőjelezhető volt.

Az eredményekből az is kiderült, hogy a modellek jelentősen pontosabb eredményekre jutottak azokban az esetekben, amelyek az emberi annotátorok számára is könnyűnek bizonyultak, vagyis ott, ahol az annotátorok elsődleges címkéi megegyeztek. Ez arra enged következtetni, hogy a nagy nyelvi modellek hasonló mintázatok és nyelvi jellegzetesség alapján sorolhatják be a szövegeket, mint az emberek. Hasonló megállapításra jutott az adathalmazon egy korábban elvégzett kutatás is a gépi tanulási módszerek kapcsán (Németh et al., 2022).

Az elsődleges és a másodlagos kategóriák sorszáma mellett a modelleket arra is megkértem, hogy adjanak egy konfidencia pontszámot 0 és 100 között arra nézve, hogy mennyire biztosak a válaszukban. Ezzel a cél annak megvizsgálása volt, hogy vajon pontosabbak-e a modellek azokban az esetekben, ahol magabiztosabbak is. Az eredmények alapján pedig az látszódott, hogy a nagyobb konfidenciájú csoportban valóban pontosabban teljesítettek a modellek, mint a közepes és alacsony értékű csoportok esetében. Ez hasznosnak bizonyulhat, ha például a nagy nyelvi modellek és a humán annotáció vegyítésén gondolkodunk, erre egy lehetőség például, ha csak azokban az esetekben alkalmazunk az emberi szupervíziót, ahol alacsonyak a modell konfidencia pontszámai.

A pontosság mellett a megbízhatóság kérdésére is kitértem a szakdolgozatomban, ezen belül egyaránt vizsgáltam, hogy különböző modellek azonos prompttal, ugyanazon modellek különböző prompttal és ugyanazon modellek ugyanazon promptokkal és beállításokkal mennyire jutnak egyetértésre egymással. Ennek a kiértékelését szintén a konzervatív és liberális kappával tettem meg. Ebben az esetben a szakirodalommal ellentétes következtetésre jutottam (Kristensen-McLachlan et al., 2023; Reiss, 2023). Természetesen a promptok vagy a modellek változtatásával az általuk adott címke valóban változott és nem volt teljesen konzisztens sem, azonban azonos promptok különböző modelleken futtatva és különböző promptok ugyanzon a modellen futtatva egyaránt nagyobb egyetértésre jutottak a címkék tekintetében, mint az eredeti korpusz humán annotátorai (Németh, Sik & Máté, 2020). Ez pedig inkább arra enged következtetni, hogy a nagy nyelvi modellek megbízhatóság szempontjából jól teljesítenek. Ezt támasztja alá az is, hogy teljesen azonos beállítások melletti újrafuttatások esetén a GPT-4o mini átlagosan 0,93-as és 0,94-es konzervatív kappát ért el, míg a Llama 3.3 70B két promptolási stratégiájánál ez az érték 0,99, amely már szinte teljes egyezésként kezelhető.

Összegezve az eredményeket a vizsgált nagy nyelvi modellek a pontosság tekintetében elfogadható módon szerepeltek, míg megbízhatóság szempontjából pedig már egészen jó eredményekre voltak képesek. A nyílt forráskódú Llama 3.3 70B jobb eredményeket ért el, mint a GPT-4o mini, habár érdemes megemlíteni, hogy ez utóbbi egy kisebb modell, amely feltehetőleg kevesebb paraméterrel rendelkezik, ugyanakkor a modell dokumentációjában erre nézve hivatalos adatot nem közölt az OpenAI. A nyílt forráskódú modellek alkalmazása azonban összeségében mindenképpen egy működőképes alternatíva lehet abban az esetben, ha az adatvédelem, a modellekre való nagyobb rálátás és az eredmények reprodukálhatósága fontos szempontként jelenik meg a kutatásunkban. Érdemes még azt is kiemelni, hogy a Groq API-n keresztül futtatott Llama 3.3 modell futásának hossza fele akkora volt, mint az OpenAI API-n futatott GPT-4o minié, ez azonban közel hétszeres költségekkel járt együtt a Llama esetében.

Összeségében a vizsgált nagy nyelvi modellek pontossága nem volt igazán kiemelkedő, ami abba az irányba mutat, hogy nem érdemes mindenféle emberi szupervízió nélkül használni őket a szövegek annotálására és kategorizálására. Ahhoz, hogy az eredményük kiértékelhető

legyen pedig mindenképpen szükséges egy legalább pár száz szövegből álló gold standard, így ezek a modellek egyelőre nem képesek a teljes humán annotáció kiváltására.

A téma átfogóbb elemzéséhez a lehetséges kutatási irányok közé tartozhat nagyobb számú és változatosabb nagy nyelvi modellek együttes vizsgálata, illetve különböző klasszifikációs és promptolási módszerek mélyebb elemzése.

Szintén érdemes lehet nagyobb hangsúlyt fektetni arra, hogy a promptok nyelvezetének vagy szövegezésének változtatása, akár csak egyazon megközelítésen belül hogyan befolyásolja a modellek teljesítményét. Ennek kapcsán hasznos lehet annak vizsgálata is, hogy a nagy nyelvi modelleknek megadott szerepkör jelentősen befolyásolja-e a kapott eredményeket. Ezt akár a prompt szövegében vagy a külön erre alkalmazott szerep (role) bemenetnél is megtehetjük, például „Te egy szociológia szakos hallgató vagy, aki depresszióval kapcsolatos posztokat címkéz” vagy „Te egy professzionális annotátor vagy”. Továbbá célszerű lenne annak elemzése is, hogy a few-shot technikánál a modelleknek adott példák mekkora hatással vannak a kapott eredményekre. Illetve jelen kutatás esetében az is egy érdekes kérdés, hogy mennyire változik a pontosság, ha eleve négy kategóriás klasszifikációt kérünk a modelltől.

További érdekes lehetőség lenne azt is megvizsgálni, hogy a különböző nagy nyelvi modellek finomhangolásával mekkora mértékben növelhető a klasszifikációk pontossága. Emellett további irányt jelenthet a magyar nyelvű szövegkorpuszok vizsgálata is, amelyek esetében elképzelhető a rosszabb teljesítmény is, ha figyelembe vesszük azt, hogy a modellek tanításánál az ilyen nyelvű szövegek jelentős kisebbségben vannak. Ezen felül érdekes irányként szolgálhat annak megvizsgálása, hogy a nagy nyelvi modellek hogyan vegyíthetők a humán annotációval, például mint azt segítő eszközök, amelyeknek célja nem az emberi címkézés kiváltása. Végül pedig az értelmezhetőség kérdésköre szintén egy olyan irány, amelyben érdemes lenne előrelépéseket tenni, akár szöveges indoklások kérésével és feldolgozásával.

Irodalomjegyzék

Aggarwal, C. C., & Zhai, C. X. (2012a). A survey of text classification algorithms. In Aggarwal, C. C., & Zhai, C. X. (Szerk.), *Mining text data* (pp. 163–222). Springer. DOI: https://doi.org/10.1007/978-1-4614-3223-4_6

Aggarwal, C. C., & Zhai, C. X. (2012b). A survey of text clustering algorithms. In Aggarwal, C. C., & Zhai, C. X. (Szerk.), *Mining text data* (pp. 77–128). Springer. DOI: https://doi.org/10.1007/978-1-4614-3223-4_4

Alizadeh, M., Kubli, M., Samei, Z., Dehghani, S., Bermeo, J. D., Korobeynikova, M., & Gilardi, F. (2023). Open-source large language models outperform crowd workers and approach ChatGPT in text-annotation tasks. *arXiv preprint arXiv:2307.02179v1*. Elérhető: https://arxiv.org/abs/2307.02179v1 (Letöltve: 2025.04.14.)

Aroyo, L., & Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine, 36*(1), 15–24. DOI: https://doi.org/10.1609/aimag.v36i1.2564

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)* (pp. 610–623). Association for Computing Machinery. DOI: https://doi.org/10.1145/3442188.3445922

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., ... Liang, P. (2021). *On the opportunities and risks of foundation models*. Center for Research on Foundation Models, Stanford University. Elérhető: https://crfm.stanford.edu/report.html (Letöltve: 2025.04.14)

Bontcheva, K., Derczynski, L., & Roberts, I. (2017). Crowdsourcing named entity recognition and entity linking corpora. In Ide, N., & Pustejovsky, J. (Szerk.), *Handbook of linguistic annotation* (pp. 875–892). Springer. DOI: https://doi.org/10.1007/978-94-024-0881-2_32

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., 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. *Advances in Neural Information Processing Systems, 33*, 1877–1901. Elérhető: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf (Letöltve: 2025.04.14)

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186. DOI: https://doi.org/10.1126/science.aal4230

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. DOI: https://doi.org/10.1177/001316446002000104

de Gibert, O., Perez, N., García-Pablos, A., & Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. In Fišer, D., Huang, R., Prabhakaran, V., Voigt, R., Waseem, Z., & Wernimont, J. (Szerk.), *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)* (pp. 11–20). Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/W18-5102

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. In Jurafsky, D., Chai, J., Schluter, N., & Tetreault, J. (Szerk.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4040–4054). Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2020.acl-main.372

Flor, M., Yoon, S. Y., Hao, J., Liu, L., & von Davier, A. (2016). Automated classification of collaborative problem solving interactions in simulated science tasks. In J. Tetreault, J. Burstein, C. Leacock, & H. Yannakoudakis (Szerk.), *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 31–41). Association for Computational Linguistics. DOI: 10.18653/v1/W16-0504

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science, 349*(6245), 261–266. DOI: https://doi.org/10.1126/science.aaa8685

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751v2*. Elérhető: https://arxiv.org/abs/1904.09751v2 (Letöltve: 2025.04.14.)

Hovy, E., & Lavid, J. (2010). Towards a 'science' of corpus annotation: A new methodological challenge for corpus linguistics. *International Journal of Translation, 22*(1), 13–36.

Geva, M., Goldberg, Y., & Berant, J. (2019). Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets. In Inui, K., Jiang, J., Ng, V., & Wan, X. (Szerk.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 1161–1166). Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/D19-1107

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences, 120*(30), e2305016120. DOI: https://doi.org/10.1073/pnas.2305016120

Goel, A., Gueta, A., Gilon, O., Liu, C., Erell, S., Nguyen, L. H., Hao, X., Jaber, B., Reddy, S., Kartha, R., Steiner, J., Laish, I., & Feder, A. (2023). LLMs accelerate annotation for medical information extraction. In Hegselmann, S., Parziale, A., Shanmugam, D., Tang, S., Asiedu, M. N., Chang, S., Hartvigsen, T., & Singh, H. (Szerk.), *Proceedings of the 3rd Machine Learning for Health Symposium* (Vol. 225, pp. 82–100). Proceedings of Machine Learning Research. Elérhető: https://proceedings.mlr.press/v225/goel23a.html (Letöltve: 2025.04.14.)

Jurafsky, D., & Martin, J. H. (2025). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (3. kiad., kézirat, 2025. január). Elérhető: https://web.stanford.edu/~jurafsky/slp3/ed3book_Jan25.pdf (Letöltve: 2025.04.14.)

Katona, E., Szeitl, B., & Túry-Angyal, E. (2025). *A kutatás ára: Az empirikus társadalomkutatás ökológiai lábnyoma és ennek mérési lehetőségei a természetesnyelv-feldolgozás példáján keresztül* [Nem publikált kézirat].

Kerekes, N. (2020). *Multi-label szövegklasszifikáció online fórumbejegyzéseken* [Nem publikált mesterszakos szakdolgozat]. Eötvös Loránd Tudományegyetem. Elérhető: https://rc2s2.elte.hu/wp-content/uploads/2021/08/KerekesNorbert_OA2ZIC_szakdolgozat_2020-21-1.pdf (Letöltve: 2025.04.14.)

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2. kiad.). Sage Publications.

Kristensen-McLachlan, R. D., Canavan, M., Kardos, M., Jacobsen, M., & Aarøe, L. (2023). Chatbots are not reliable text annotators. *arXiv preprint arXiv:2311.05769v1*. Elérhető: https://arxiv.org/abs/2311.05769v1 (Letöltve: 2025.04.14.)

Labruna, T., Brenna, S., Zaninello, A., & Magnini, B. (2023). Unraveling ChatGPT: A critical analysis of AI-generated goal-oriented dialogues and annotations. In Basili, R., Lembo, D., Limongelli, C., & Orlandini, A. (Szerk.), *AIxIA 2023 – Advances in Artificial Intelligence: XXIInd International Conference of the Italian Association for Artificial Intelligence, AIxIA 2023, Rome, Italy, November 6–9, 2023, Proceedings*. *Lecture Notes in Computer Science* (Vol. 14318, pp. 151–171). Springer. DOI: https://doi.org/10.1007/978-3-031-47546-7_11

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. DOI: https://doi.org/10.1038/nature14539

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781.* Elérhető: https://arxiv.org/abs/1301.3781 (Letöltve: 2025.04.14.)

Németh, R. (2024). *Az automatizált szövegfeldolgozás szociológiai lehetőségei*. Savaria University Press.

Németh, R., Katona, E. R., & Kmetty, Z. (2020). Az automatizált szövegelemzés perspektívája a társadalomtudományokban. *Szociológiai Szemle, 30*(1), 44–62. DOI: https://doi.org/10.51624/SzocSzemle.2020.1.3

Németh, R., Máté, F., Katona, E., Rakovics, M., & Sik, D. (2022). Bio, psycho, or social: Supervised machine learning to classify discursive framing of depression in online health communities. *Quality & Quantity, 56*, 3933–3955. DOI: https://doi.org/10.1007/s11135-021-01299-0

Németh, R., Sik, D., & Máté, F. (2020). Machine learning of concepts hard even for humans: The case of online depression forums. *International Journal of Qualitative Methods, 19*, 1–8. DOI: https://doi.org/10.1177/1609406920949338

Ollion, É., Shen, R., Macanovic, A., & Chatelain, A. (2023). ChatGPT for text annotation? Mind the hype! *SocArXiv preprint*. Elérhető: https://osf.io/preprints/socarxiv/x58kn_v1 (Letöltve: 2025.04.14.)

Ostyakova, L., Smilga, V., Petukhova, K., Molchanova, M., & Kornev, D. (2023). ChatGPT vs. crowdsourcing vs. experts: Annotating open-domain conversations with speech functions. In Stoyanchev, S., Joty, S., Schlangen, D., Dušek, O., Kennington, C., & Alikhani, M. (Szerk.), *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 242–254). Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2023.sigdial-1.23

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*, 27730–27744.

Qureshi, K. A., & Sabih, M. (2021). Un-compromised credibility: Social media based multi-class hate speech classification for text. *IEEE Access, 9*, 109465–109477. DOI: https://doi.org/10.1109/ACCESS.2021.3101977

Reiss, M. V. (2023). Testing the reliability of ChatGPT for text annotation and classification: A cautionary remark. *arXiv preprint arXiv:2304.11085v1*. Elérhető: https://arxiv.org/abs/2304.11085v1 (Letöltve: 2025.04.14.)

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. In Korhonen, A., Traum, D., & Màrquez, L. (Szerk.), *Proceedings of the*

57th Annual Meeting of the Association for Computational Linguistics (pp. 1668–1678). Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/P19-1163

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. (2008). Cheap and fast – but is it good? Evaluating non-expert annotations for natural language tasks. In Lapata, M., & Ng, H. T. (Szerk.), *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 254–263). Association for Computational Linguistics.

Törnberg, P. (2023). ChatGPT-4 outperforms experts and crowd workers in annotating political Twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588.* Elérhető: https://arxiv.org/abs/2304.06588 (Letöltve: 2025.04.14.)

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008. Elérhető: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (Letöltve: 2025.04.14.)

Wahba, G. (2002). Soft and hard classification by reproducing kernel Hilbert space methods. *Proceedings of the National Academy of Sciences, 99*(26), 16524–16530. DOI: https://doi.org/10.1073/pnas.242574899

Zhu, Y., Zhang, P., Haq, E.-U., Hui, P., & Tyson, G. (2023). Can ChatGPT reproduce human-generated labels? A study of social computing tasks. *arXiv preprint arXiv:2304.10145*. Elérhető: https://arxiv.org/abs/2304.10145v2 (Letöltve: 2025.04.14.)

Függelék

1. számú függelék – Alkalmazott promptok

a. Rövid zero-shot prompt (kemény klasszifikáció)

Az alábbi promptot a kemény klasszifikáció rövid zero-shot változataként használtam, ahol a modell csak az alapvető utasításokat és a kategóriák listáját kapta meg:

Read the following post about depression and classify it based on how it frames the topic. The framing of depression refers to how patients perceive and interpret their illness. Framing determines the causal explanations patients give for their illness and the treatment they choose to recover from it. Provide a primary category, and if applicable, a secondary category. If it is unclear whether it fits into biomedical, psychological, or sociological, classify it as not relevant or unclassifiable. Also, provide a single confidence score between 0 and 100 for your classification as a whole.

1: biomedical

2: psychological

3: sociological

4: unclassifiable

5: irrelevant

Respond with the primary category number (1-5), optionally followed by a secondary category number (1-3 or 999 if there is no secondary category), and then a single confidence score (0-100). Follow this output example: '2 3 85'. Do not write anything else, besides the three numbers.

Post: {text}

b. Részletes zero-shot prompt (kemény klasszifikáció)

Ebben a változatban a modell már részletes definíciót is kapott az egyes kategóriákról:

The framing of depression refers to how patients perceive and interpret their illness. Framing determines the causal explanations patients give for their illness and the treatment they choose to recover from it. In scientific discourses, three main framings dominate the debate on depression: A biomedical one (referring to depression as a disease of physical origin, dysfunction or illness, requiring medical intervention), a psychological one (referring to depression as a result of cognitive, affective or behavioural dysfunction, result of maladaptation/trauma, requiring psychotherapy), and a sociological one (referring to depression as a sociological phenomenon, consequence of social distortions caused by

peripheral role experiences in the social network, requiring social work/social policy intervention). We collected depression-related posts from the most popular English-speaking health forums. I'm going to give you a random sample of them. I ask you to sort them into 5 categories (labelled 1, 2, 3, 4 or 5), where label 1 represents biomedical framing, 2 represents psychological framing, 3 represents sociological framing, 4 if you cannot classify the framing (it is about depression but the framing cannot be identified) and 5 if the post is irrelevant (it is not about depression).

You can give two labels if more than one interpretation is possible for the post. The primary label reflects what you consider to be the main type of framing in the given text, and a secondary label reflects an additional type of framing.

On a scale of 0–100, please also indicate how confident you are in your assessment of how clear the categorisation of the post was.

So for each post we expect 3 responses: primary label (1–5), secondary label (1–3 or 999 if there is no secondary category) and confidence in answer (0–100). Follow this output example: '2 3 85'. You should never use four as a secondary label. You must always provide all three values. Do not write anything else, besides the three numbers.

Post: {text}

c. Few-shot prompt (kemény klasszifikáció)

Az alábbi prompt a few-shot technikát alkalmazza, vagyis a modell nemcsak a kategóriák meghatározását, hanem minden kategóriára több konkrét példát is kapott, hogy az osztályozási szempontokat jobban megértse:

Read the following post about depression and classify it based on how it frames the topic. The framing of depression refers to how patients perceive and interpret their illness. Framing determines the causal explanations patients give for their illness and the treatment they choose to recover from it. Provide a primary category, and if applicable, a secondary category. If it is unclear whether it fits into biomedical, psychological, or sociological, classify it as not relevant or unclassifiable. Also, provide a single confidence score between 0 and 100 for your classification as a whole.

Categories:

1: biomedical – referring to depression as a disease of physical origin, dysfunction or illness, requiring medical intervention

2: psychological – referring to depression as a result of cognitive, affective or behavioural dysfunction, result of maladaptation/trauma, requiring psychotherapy

3: sociological – referring to depression as a sociological phenomenon, consequence of social distortions caused by peripheral role experiences in the social network, requiring social work/social policy intervention

4: unclassifiable – it is about depression but the framing cannot be identified

5: irrelevant – it is not about depression

Examples:

Post: I'm sorry .. I shouldn't have answered your question with a long reply like that .. I was in a bad mood yesterday & lost it .. (Category: 5 - irrelevant)

Post: Welcome to the forum my friend and I hate to hear that you are having some difficult times right now. Hang in there my friend and continue to keep coming back to this forum and you might can learn some different ways of coping with your mental health issues (Category: 5 - irrelevant)

Post: Our experiences of Churches has been rather different but with the same result - we stopped going. The sermons were too long, and boring, the congregation seemed mostly asleep and no-one was invited to participate, if anyone wanted to anyway, and we experienced little or no fellowship. But enough of this, my wife and myself will continue to pray, ask God for help to endure, guidance, and protect us from the wickedness of the world. (Category: 4 - unclassifiable)

Post: I call them "episode" when they become much worse to the point I can't function anymore. On normal days I deal with "residual" symptoms (or mild ones). (Category: 4 - unclassifiable)

Post: What do you find helps you more? I took just an AP for years and was fairly stable. I feel even more stable with a mood stabilizer. I have heard of people taking just a mood stabilizer too. Also, which one has more severe side effects? (Category: 1 - biomedical)

Post: I've ruined my life from drug use. I'm 7 months sober and initially things were going pretty great in my recovery. I was playing music (guitar , piano, banjo) a ton but my forearm tendonitis has gotten so fucking bad I just quit. I've seen doctors , physical therapists and had steriod injections. All to no avail. All I want to do is play and I just can't anymore. When I first got sober skateboarding saved my life but the I tore my acl and had surgery. Now I can't do any of the usual physical activity I did like yoga , weight lifting , snowboarding /skateboarding. I worked so hard over 7 months to develop these hobbies and they worked pretty damn well at keeping me moving forward. I spent 10 years shooting meth and doing nothing. It feels like matter what I do I'll never get out this hole. I owe everyone money, I have a shit job, and I'm just fucking done. Life doesn't feel worth it. I've got two cats that I love to death. If it weren't for them I would peace the fuck out of this small town and get some dope. Ive been trying desperately to find dope but this town is to small. (Category: 1 - biomedical, 2 - psychological)

Post: Today I have noticed... That being true, allowing my emotion to be free, allows me to be a good friend to myself. I am not good at emotional regulation. My emotion is out of control. My feelings from the past rage on. That is not accepted in this world. I am not hurtful to others but myself I can be cruel.... Who can accept me AS IS... I wish I knew how to get full grasp of my feelings.... (Category: 2 - psychological, 3 - sociological)

Post: I know what You mean, but it is risky. Almost every time I went outside my comfort zone I ended up with self-inflicted wounds. Its just too dangerous. I did volunteer work before. I even traveled to a different city alone to try changing it when I was only 16. All of this changed nothing. I really like your point of view. Unfortunatly it does not work in my case. (Category: 2 - psychological)

Post: If you don't want to go don't go. I gave my heart my soul to my work for years. But I have realised at the end if you are healthy to do their job you are best. If some little thing goes wrong with your health they don't want to know you. And to be honest the people you work with drains ruins your life. I have worked 17 years for NHS. After 2010 all management changed and my life become harder and harder to cope with new management and at the end I could not go back because of my anxiety and depression. If you could find another job before it comes to the stage like mine please look for a another job.. You should always thing

of your self first not them.. Don't allow them to drain you and your family.. I still feel destroyed by my work place.. ALL the best my love xxxx (Category: 3 - sociological)

Post: i can relate to this.the only thing is i worked in dietary for a hospital should be caring right? wrong. i worked there for 10 years,never called out was always on time and gave 110% to my job. i broke my foot in october and had to go on medical leave; they hole your jobs for 3months. i needed one more week before i could return i appled for an extension and was denied and lost my job. the moral is no one really cares about you or your problems and the job was so stressful i hated to go in because management kept dumping more work on us. so that added to my anxiety and depression. i even had tmj which is a jaw problem. since i'm out of there a lot of my symptoms have lessened. so i can totally relate to you (Category: 3 - sociological, 1 - biomedical)

Respond with the primary category number (1-5), optionally followed by a secondary category number (1-3 or 999 if there is no secondary category), and then a single confidence score (0-100). Follow this output example: '2 3 85'. You should never use four as a secondary category. You must always provide all three values. Do not write anything else, besides the three numbers.

Post: {text}

d. Rövid zero-shot prompt (puha klasszifikáció)

A korábbi zero-shot prompt módosított verziója, amely a puha klasszifikációt alkalmazza:

Read the following post about depression and classify it based on how it frames the topic. The framing of depression refers to how patients perceive and interpret their illness. Framing determines the causal explanations patients give for their illness and the treatment they choose to recover from it.

Assign a probability between 0 and 1 to each of the following five categories:

1: biomedical

2: psychological

3: sociological

4: unclassifiable

5: irrelevant

Also provide an overall confidence score between 0 and 100, indicating how confident you are in your classification.

Respond with five decimal numbers (0–1) and one integer (0–100), space-separated, in the following order: score_biomedical score_psychological score_sociological score_unclassifiable score_irrelevant confidence_score

Follow this output example exactly: '0.65 0.25 0.08 0.01 0.01 85'. Do not write anything else.If you're uncertain, estimate the most plausible category rather than assigning all zeros.

When assigning probabilities, ensure that each non-zero value is slightly different from the others. Only assign the same value to multiple categories if that value is 0.00.

Post: {text}

e. Részletes zero-shot prompt (puha klasszifikáció)

A korábbi hosszabb zero-shot prompt módosított verziója, amely a puha klasszifikációt alkalmazza:

The framing of depression refers to how patients perceive and interpret their illness. Framing determines the causal explanations patients give for their illness and the treatment they choose to recover from it. In scientific discourses, three main framings dominate the debate on depression: A biomedical one (referring to depression as a disease of physical origin, dysfunction or illness, requiring medical intervention), a psychological one (referring to depression as a result of cognitive, affective or behavioural dysfunction, result of maladaptation/trauma, requiring psychotherapy), and a sociological one (referring to depression as a sociological phenomenon, consequence of social distortions caused by peripheral role experiences in the social network, requiring social work/social policy intervention). We collected depression-related posts from the most popular English-speaking health forums. I'm going to give you a random sample of them. I ask you to assign probabilities (between 0 and 1) to each of the following five categories: 1 for biomedical framing, 2 for psychological framing, 3 for sociological framing, 4 if you cannot classify the framing (unclassifiable), and 5 if the post is irrelevant (not about depression).

So for each post we expect 6 responses: five probabilities (0–1) in order and one confidence score (0–100). Follow this output example exactly: '0.65 0.25 0.08 0.01 0.01 85'. Do not write anything else.If you're uncertain, estimate the most plausible category rather than assigning all zeros.

When assigning probabilities, ensure that each non-zero value is slightly different from the others. Only assign the same value to multiple categories if that value is 0.00.

Post: {text}

f. Few-shot prompt (puha klasszifikáció)

A few-shot prompt módosított verziója, amely a puha klasszifikációt alkalmazza:

Read the following post about depression and classify it based on how it frames the topic. The framing of depression refers to how patients perceive and interpret their illness. Framing determines the causal explanations patients give for their illness and the treatment they choose to recover from it.

Assign a probability between 0 and 1 to each of the following five categories:

Categories:

1: biomedical – referring to depression as a disease of physical origin, dysfunction or illness, requiring medical intervention

2: psychological – referring to depression as a result of cognitive, affective or behavioural dysfunction, result of maladaptation/trauma, requiring psychotherapy

3: sociological – referring to depression as a sociological phenomenon, consequence of social distortions caused by peripheral role experiences in the social network, requiring social work/social policy intervention

4: unclassifiable – it is about depression but the framing cannot be identified

5: irrelevant – it is not about depression

Examples:

Post: I'm sorry .. I shouldn't have answered your question with a long reply like that .. I was in a bad mood yesterday & lost it .. (Category: 5 - irrelevant)

Post: Welcome to the forum my friend and I hate to hear that you are having some difficult times right now. Hang in there my friend and continue to keep coming back to this forum and you might can learn some different ways of coping with your mental health issues (Category: 5 - irrelevant)

Post: Our experiences of Churches has been rather different but with the same result - we stopped going. The sermons were too long, and boring, the congregation seemed mostly asleep and no-one was invited to participate, if anyone wanted to anyway, and we experienced little or no fellowship. But enough of this, my wife and myself will continue to pray, ask God for help to endure, guidance, and protect us from the wickedness of the world. (Category: 4 - unclassifiable)

Post: I call them "episode" when they become much worse to the point I can't function anymore. On normal days I deal with "residual" symptoms (or mild ones). (Category: 4 - unclassifiable)

Post: What do you find helps you more? I took just an AP for years and was fairly stable. I feel even more stable with a mood stabilizer. I have heard of people taking just a mood stabilizer too. Also, which one has more severe side effects? (Category: 1 - biomedical)

Post: I've ruined my life from drug use. I'm 7 months sober and initially things were going pretty great in my recovery. I was playing music (guitar , piano, banjo) a ton but my forearm tendonitis has gotten so fucking bad I just quit. I've seen doctors , physical therapists and had steriod injections. All to no avail. All I want to do is play and I just can't anymore. When I first got sober skateboarding saved my life but the I tore my acl and had surgery. Now I can't do any of the usual physical activity I did like yoga , weight lifting , snowboarding /skateboarding. I worked so hard over 7 months to develop these hobbies and they worked pretty damn well at keeping me moving forward. I spent 10 years shooting meth and doing nothing. It feels like matter what I do I'll never get out this hole. I owe everyone money, I have a shit job, and I'm just fucking done. Life doesn't feel worth it. I've got two cats that I love to death. If it weren't for them I would peace the fuck out of this small town and get some dope. Ive been trying desperately to find dope but this town is to small. (Category: 1 - biomedical, 2 - psychological)

Post: Today I have noticed... That being true, allowing my emotion to be free, allows me to be a good friend to myself. I am not good at emotional regulation. My emotion is out of control. My feelings from the past rage on. That is not accepted in this world. I am not hurtful to others but myself I can be cruel.... Who can accept me AS IS... I wish I knew how to get full grasp of my feelings.... (Category: 2 - psychological, 3 - sociological)

Post: I know what You mean, but it is risky. Almost every time I went outside my comfort zone I ended up with self-inflicted wounds. Its just too dangerous. I did volunteer work before. I even traveled to a different city alone to try changing it when I was only 16. All of this changed nothing. I really like your point of view. Unfortunatly it does not work in my case. (Category: 2 - psychological)

Post: If you don't want to go don't go. I gave my heart my soul to my work for years. But I have realised at the end if you are healthy to do their job you are best. If some little thing goes wrong with your health they don't want to know you. And to be honest the people you work with drains ruins your life. I have worked 17 years for NHS. After 2010 all management changed and my life become harder and harder to cope with new management and at the end I could not go back because of my anxiety and depression. If you could find another job before it comes to the stage like mine please look for a another job.. You should always thing of your self first not them.. Don't allow them to drain you and your family.. I still feel destroyed by my work place.. ALL the best my love xxxx (Category: 3 - sociological)

Post: i can relate to this.the only thing is i worked in dietary for a hospital should be caring right? wrong. i worked there for 10 years,never called out was always on time and gave 110% to my job. i broke my foot in october and had to go on medical leave; they hole your jobs for 3months. i needed one more week before i could return i appled for an extension and was denied and lost my job. the moral is no one really cares about you or your problems and the job was so stressful i hated to go in because management kept dumping more work on us. so that added to my anxiety and depression. i even had tmj which is a jaw problem. since i'm out of there a lot of my symptoms have lessened. so i can totally relate to you (Category: 3 - sociological, 1 - biomedical)

Also provide an overall confidence score between 0 and 100, indicating how confident you are in your classification.

Respond with five decimal numbers (0–1) and one integer (0–100), space-separated, in the following order:

score_biomedical score_psychological score_sociological score_unclassifiable score_irrelevant confidence_score

Follow this output example exactly: '0.65 0.25 0.08 0.01 0.01 85'. Do not write anything else.If you're uncertain, estimate the most plausible category rather than assigning all zeros.

When assigning probabilities, ensure that each non-zero value is slightly different from the others.Only assign the same value to multiple categories if that value is 0.00.

Post: {text}

2. számú függelék – A tanítóhalmazon futtatott modellek kappa eredményei különböző temperature beállítások mellett

12. táblázat GPT-4o mini, rövid zero-shot prompt (kemény klasszifikáció)

	Temp. = 0	Temp. = 0,2	Temp. = 0,4	Temp. = 0,6
Konzervatív kappa	0,34	0,34	0,33	0,34
Liberális kappa	0,44	0,43	0,43	0,43

13. táblázat GPT-4o mini, részletes zero-shot prompt (kemény klasszifikáció)

	Temp. = 0	Temp. = 0,2	Temp. = 0,4	Temp. = 0,6
Konzervatív kappa	0,43	0,43	0,43	0,43
Liberális kappa	0,51	0,51	0,52	0,51

14. táblázat GPT-4o mini, few-shot prompt (kemény klasszifikáció)

	Temp. = 0	Temp. = 0,2	Temp. = 0,4	Temp. = 0,6
Konzervatív kappa	0,47	0,47	0,46	0,46
Liberális kappa	0,55	0,55	0,55	0,54

15. táblázat GPT-4o mini, konzervatív kappa a különböző puha klasszifikációs promptolási technikáknál

	Temp. = 0	Temp. = 0,2	Temp. = 0,4	Temp. = 0,6
Rövid zero-shot	0,22	0,22	0,22	0,22

Részletes zero-shot	0,38	0,37	0,38	0,38
Few-shot	0,43	0,42	0,41	0,41

16. táblázat Llama 3.3, rövid zero-shot prompt (kemény klasszifikáció)

	Temp. = 0	**Temp. = 0,2**	**Temp. = 0,4**	**Temp. = 0,6**
Konzervatív kappa	0,38	0,39	0,38	0,38
Liberális kappa	0,48	0,48	0,48	0,48

17. táblázat Llama 3.3, részletes zero-shot prompt (kemény klasszifikáció)

	Temp. = 0	**Temp. = 0,2**	**Temp. = 0,4**	**Temp. = 0,6**
Konzervatív kappa	0,46	0,46	0,47	0,46
Liberális kappa	0,57	0,57	0,58	0,57

18. táblázat Llama 3.3, few-shot prompt (kemény klasszifikáció)

	Temp. = 0	**Temp. = 0,2**	**Temp. = 0,4**	**Temp. = 0,6**
Konzervatív kappa	0,49	0,48	0,49	0,48
Liberális kappa	0,58	0,58	0,58	0,57

19. táblázat Llama 3.3, konzervatív kappa a különböző puha klasszifikációs promptolási technikáknál

	Temp. = 0	**Temp. = 0,2**	**Temp. = 0,4**	**Temp. = 0,6**
Rövid zero-shot	0,32	0,33	0,32	0,33
Részletes zero-shot	0,39	0,39	0,39	0,39
Few-shot	0,45	0,45	0,44	0,43

3. számú függelék – GPT-4o mini few-shot confusion mátrix

9. ábra A valós elsődleges és a modell által adott elsődleges címkék confusion mátrixa