Melyik módszert válasszuk a jegybanki kommunikáció szövegelemzésére? Mesterséges intelligencia és korábbi technikák összevetése*

Kocsis Zalán – Mátrai-Pitz Mónika

Tanulmányunk szövegelemzési módszerek tulajdonságait hasonlítja össze az amerikai Federal Reserve és négy kelet-közép-európai jegybank kommunikációinak szöveges mintáin. Eredményeink alapján a jegybanki szövegekben rejlő monetáris politikai, reálgazdasági és inflációs információt a BERT-típusú modelleken alapuló módszerek tudják legpontosabban megragadni, megelőzve az OpenAI GPT-4.1 és GPT-5 modelljeit. A BERT-típusú módszerek a GPT-modelleknél gyorsabbak, offline futtathatók előfizetés nélkül, hátrányuk azonban, hogy a modellek külön tanítást igényelnek, ami hardver- és munkaerőköltséget jelent, és a módszer változtatásait is nehézkesebbé teszi. A GPT-modellek ellenben rugalmasabbak és új jegybanki mintákon pontosabbaknak bizonyultak. A benchmarkként alkalmazott szótáralapú módszerek nagyságrendekkel pontatlanabbak, használatukat ugyanakkor bizonyos esetekben a gyorsaság, a költségmentes futtatás, illetve a módszer transzparenciája indokolhatja.

Journal of Economic Literature (JEL) kódok: C55, C63, E52, E58

Kulcsszavak: jegybanki közlemények, mesterséges intelligencia, szövegelemzés

1. Bevezetés

A jegybanki kommunikáció gépi szövegelemzése mind elterjedtebbé válik, ahogy ezek a módszerek egyre pontosabbá válnak. Tanulmányunk elsődleges célja arra a természetesen felmerülő kérdésre választ adni, hogy a sok létező módszer közül milyen körülmények között, melyiket érdemes használni. Ennek érdekében a jelenleg kurrens legnépszerűbb szövegelemző módszereket több tulajdonság mentén (pontosság-, gyorsaság-, költségszempontok) hasonlítjuk össze. A módszerek

* A jelen kiadványban megjelenő írások a szerzők nézeteit tartalmazzák, ami nem feltétlenül egyezik a Magyar Nemzeti Bank hivatalos álláspontjával.

Kocsis Zalán: Magyar Nemzeti Bank, vezető közgazdasági szakértő. E-mail: kocsisz@mnb.hu
Mátrai-Pitz Mónika: Magyar Nemzeti Bank, vezető adattudós. E-mail: pitzm@mnb.hu

Köszönettel tartozunk Kovács Péter Márk, Lang Adrienn, Molnár Tamás, Nagy Balázs, Pásztor Szabolcs, Simon Péter kollégáinknak, akik segítő észrevételeikkel, illetve a kézi annotációk létrehozásával hozzájárultak a tanulmány elkészüléséhez. A fennmaradó hibákért a felelősség kizárólag a szerzőket terheli.

A magyar nyelvű kézirat első változata 2025. szeptember 15-én érkezett szerkesztőségünkbe.

DOI: https://doi.org/10.25201/HSZ.24.4.34

hatékonysága közötti különbségtétel érdekes lehet a jegybanki kommunikációval foglalkozó szakirodalom szempontjából – például hogy mely módszereket érdemes használni a kommunikáció piaci hatásainak mérésében –, vagy, hogy melyik módszerrel lehet pontosabban interpretálni a jegybanki kommunikációt a piaci szereplők számára, illetve a jegybankok számára is, akik a saját hatásukat, illetve a piaci szereplők információit igyekeznek monitorozni.

A jegybankok kommunikációja a kilencvenes évek vége óta egyre fontosabb szerepet töltött be a monetáris politikai eszköztárban. A döntések hátterét ismertető kommunikáció, illetve a jegybanki előrejelzéseket nyilvánosságra hozó előretekintő iránymutatás (forward guidance) több célt szolgál.[1] Egyrészt a jegybanki transzparencia a gazdasági szereplők várakozásainak horgonyzásán keresztül segít elérni a jegybanki inflációs célt, és mérsékelheti a rövid távú sokkok beépülését a hosszú távú várakozásokba (*Blinder et al. 2008*; *Dovern et al. 2012; Eusepi – Preston 2010*). Másrészt a jegybank előrejelzéseinek és reakciófüggvényének a pontosabb ismerete összehangolja a piaci szereplők várakozásait (*Ehrmann et al. 2010*; *Naszódi et al. 2016; Seelajaroen 2019*), ami növeli a gazdaság kiszámíthatóságát, csökkentve a piaci volatilitást és a hozamok lejárati prémium tartalmát (*Poole et al. 2002*; *Gábriel – Pintér 2006*; *Horváth et al. 2014*). Harmadrészt a kommunikációval a jegybank a hagyományos alapkamat döntésekhez képest a piaci hozamgörbe szélesebb és a reálgazdaság szempontjából relevánsabb horizont spektrumait tudja elérni, ami segíti a monetáris transzmissziót.[2]

A jegybanki kommunikáció minél pontosabb interpretációja fontos a piaci szereplőknek. A kommunikáció egyrészt a jegybanki reakciófüggvénnyel kapcsolatos információt közvetíti, ami alapján következtethető, hogy a jegybank milyen változókat milyen súllyal vesz figyelembe a jövőbeli döntéseinél. Másrészt a kommunikáció a gazdaság állapotára vonatkozó jegybanki nézeteket és előrejelzéseket is tartalmazza (az ún. információs hatást, *Romer – Romer 2000*). Az információs hatás azért lehet hasznos a piaci szereplők számára, mert a jegybank relatíve több erőforrást tud az előrejelzésekre fordítani (pl. nagyobb és felkészültebb szakmai stábot, mint a legtöbb piaci szereplő). A kommunikáció információtartalmának pontosabb interpretációja a piaci szereplők számára értékes, mivel a pontosabb előrejelzések versenyelőnyt jelenthetnek (pl. a többi szereplőnél pontosabb kamatpálya-előrejelzésnek megfelelő kereskedés várhatóan nyereséges stratégia), másrészt általában is megalapozottabbá tehetik befektetési döntéseiket.

[1] A jegybanki kommunikáció korai nemzetközi szakirodalmának átfogó összegzését adja *Blinder et al.* (*2008*). A hazai szakirodalomból kiemelnénk az előretekintő iránymutatás tapasztalataival foglalkozó *Csortos et al.* (*2014*), *Bihari* (*2015*) és *Bihari – Sztanó* (*2015*), fejlett és feltörekvő piaci jegybanki kommunikációs gyakorlatot összevető *Nagy Mohácsi et al.* (*2024*) tanulmányokat.

[2] *Gürkaynak et al.* (*2005*) empirikus elemzése bemutatta, hogy a kommunikáció piaci hatása a Fed kamatdöntésének napjain nagyobb volt a hagyományos kamatpolitikai döntések hatásainál már az általuk elemzett, főleg kilencvenes éveket tartalmazó, időbeli mintán is. Később, a 2010-es években, amikor a hagyományos kamatpolitikai eszköz a nulla alsó korlát (*zero lower bound*) elérésével lényegesen veszített a jelentőségéből, a jegybanki kommunikációs eszköz még inkább felértékelődött.

Tanulmányunk a jegybanki kommunikáció szakirodalma mellett az automatikus szövegelemzés informatikai szakirodalmához kapcsolódik[3], amelynek módszerei a közgazdasági szakirodalomba is átgyűrűztek.[4] A szövegelemzés közgazdasági irodalmának számunkra leginkább releváns ága kifejezetten jegybanki kommunikáció elemzésével foglalkozik. A jegybanki tonalitást[5] eredetileg különböző bonyolultságú szótár/szabály alapú megközelítések (*Apel – Blix Grimaldi 2012*; *Hansen – McMahon 2016*; *Correa et al. 2017*) dominálták, a 2010-es évtized közepétől ehhez egyszerűbb felügyelt tanítási algoritmusok (*Tobback et al. 2017*) és a topik-azonosításhoz nem felügyelt tanítási algoritmusok csatlakoztak (*Hansen – McMahon 2016; Hansen et al. 2018; Thorsrud 2018*), majd a 2020-as évektől a BERT-típusú nagy nyelvi modellek alkalmazásai lettek egyre népszerűbbek[6] (*Pfeifer – Marohl 2023*; *Gambacorta et al. 2024*). Az elmúlt 2–3 évben a generatív mesterséges intelligencia modelljeinek felhasználása is megkezdődött (*Fanta – Horváth 2024*; *Hansen – Kazinnik 2024; Peskoff et al. 2024; Geiger et al. 2025*).

Tanulmányunk célkitűzéséből adódóan leginkább ahhoz a szegmenshez kapcsolódik, amely szintén szövegelemzési módszerek összevetésével foglalkozik. A mi tanulmányunk konkrétabban és elsődlegesen a módszereknek azon képességét igyekszik felmérni, hogy azok mennyire hatékonyak három fundamentum (monetáris politika, reálgazdazdasági aktivitás, infláció) azonosításában, és azok tonalitásának megítélésében (szigorú/laza a monetáris politika és magas/alacsony, gyorsuló/lassuló a másik kettő esetében).

[3] A gépi szövegelemzés kezdetben igyekezett a hagyományos nyelvészeti elemzések lépéseit automatizálni, így eredetileg szabályalapú megközelítést követett. A jelenleg használt modellekhez vezető út azonban ettől az iránytól eltávolodott, és a felügyelt (tehát humán annotált mintát is felhasználó) és felügyelet nélküli gépi tanulás irányába vezetett. Ezen az úton kiemelkedő állomás volt a (neurális hálókon alapuló) mélytanulási algoritmusok alkalmazása a szövegek szavainak (tokenjeinek) numerikussá alakításában (Word2Vec: *Mikolov et al. 2013*, GloVe: *Pennington et al. 2014*), majd a transzformer architektúra (*Vaswani et al. 2017*), amely a szövegkörnyezetet is jobban képes volt megragadni. A transzformereken alapuló nagy nyelvi modellek közül az egyik első igazán népszerű és ma is sokat használt modell a BERT – Bi-directional Encoder Representation from Transformers (*Devlin et al. 2019*), amely belső struktúrája és paraméterbecslési módszere lehetővé tette, hogy humán annotációk helyett öntanuló módon eljusson egy olyan alapmodellig (ez az ún. előtanulási fázis), amely már önmagában is fejlett szövegelemző képességekkel bír. A BERT-modellek fennmaradó népszerűségének oka, hogy forráskódja és az alapmodell paraméterei bárki által elérhetőek, így a modell alkalmazásához nincs szükség az igazán erőforrásigényes előtanításra, elég csak a feladat-specifikus alkalmazásra adaptálni a modellt, az ún. finomhangolással (amihez viszont jellemzően még mindig kell humán annotáció). A 2020-as évek fejlődése egyrészt az egyre nagyobb teljesítményű és méretű nyelvi modellek, másrészt a generatív modellek (pl. GPT: *Brown et al. 2020*, Llama: *Touvron et al. 2023*) térhódításáról szólt, amelyek a feladat-specifikus felügyelt tanítást is egyre inkább ki tudták váltani. Jelenleg a legfontosabbnak tűnő fejlődési irányok egyrészt a generatív modellek eszközhasználatáról szólnak, ennek részei a RAG-alkalmazások (*Lewis et al. 2020*), amely a háttérdokumentumtárral való hatékony interakciókat jelenti; a kódírási és webkeresési képesség; illetve az általánosabb önálló eszközhasználat és azzal kapcsolatos optimalizáció ReACT (*Yao et al. 2023*), amelyre a mesterségesintelligencia-ügynökök, ügynöki rendszerek épülnek. Egy másik fejlődési irányt a modellek költség- és mérethatékonyabbá tétele jelöli ki (pl. DeepSeek: *Liu et al. 2024*).

[4] *Gentzkow et al.* (*2019*) részletes összefoglalását adja ennek a folyamatnak a 2010-es évek végéig.

[5] Tonalitáson az irány (előjel) és az intenzitás szorzatát értjük, például egy monetáris politikával kapcsolatos mondat lehet kisebb-nagyobb mértékben szigorú vagy laza irányultságú.

[6] Ide értjük a szakterületünkön népszerű *Liu et al.* (*2019*) RoBERTa (Robustly Optimized BERT Approach) és *Huang et al.* (*2022*) FinBERT modellekre épülő további modelleket is, amelyek szintén az eredeti BERT-tanulmányban (*Devlin et al. 2019*) bemutatott architektúrán alapulnak.

Tanulmányunkhoz legközelebb *Kim et al.* (*2024*), *Gambacorta et al.* (*2024*), *Pfeifer – Marohl* (*2023*) *és Hansen – Kazinnik* (*2024*) munkái állnak, amelyek szintén jegybanki szövegek elemzésének különböző módszereivel foglalkoznak és azok összehasonlítását is adják, köztük több, általunk is használt BERT-típusú módszerrel és – egy szériával régebbi – GPT-modellekkel. Az egyik fő különbség ezekhez a tanulmányokhoz képest, hogy esetünkben több ország jegybanki szövegein mérjük a modellek teljesítményét, ezáltal arra a kérdésre is választ adhatunk, hogy egyes modellek mennyire tudnak általánosan jól teljesíteni különböző jegybankok szövegmintáin. A másik fontos különbség, hogy tanulmányunk mondatonként három fundamentumra külön-külön tartalmaz humán annotációkat[7], míg ezek a tanulmányok tipikusan vagy a mondatok általános szentimentjét (pozitív/semleges/negatív) vagy a mondatok monetáris politikai irányultságának indikációit fogalmazzák meg (a laza monetáris politikától a szigorú monetáris politikáig) topik-bontás nélkül. A topik-bontás értelemszerűen a módszerek fundamentumonként eltérő teljesítményére is rámutat.

A tanulmányunkhoz kifejlesztett GPT-promptok a jegybanki kommunikáció tonalitáson túlmutató információinak (részletesebb topikok, szint/változás, időbeliség, tény/várakozás) kinyerése irányába is tesz lépéseket, ilyen szempontból kapcsolódunk *Byrne et al.* (*2023*), *Evdokimova et al.* (*2023*), *Geiger et al.* (*2025*) és *Yao et al.* (*2025*) munkáihoz, amelyek ezek egyes aspektusaival szintén foglalkoznak.

Annotációs adatbázisunkat és kódjaink jelentős részét publikáljuk, hogy ezáltal is segítsük a szakterület kutatási/elemzési tevékenységét.[8]

2. Adatok és módszerek

2.1. Közleményminta és humán annotációk

Közlemény adataink három forrásból származnak, melyek közül a Magyar Nemzeti Bank (MNB) közleményei adják a legnagyobb részt. A 2017–2024 időszakból véletlenszerűen választott 32 közlemény szövegei összesen 2 465 mondatot tesznek ki. Másrészt két tanulmány (*Gorodnichenko et al. 2023*; *Niţoi et al. 2023*) által közzétett, az interneten szabadon elérhető szövegeket használtuk, illetve azok annotációiból

[7] Humán annotációk (másnéven manuális vagy kézi címkék) a manuálisan elvégzett mondat-kiértékeléseket jelentik. Ezeknek az annotációknak a célja, hogy „javítókulcsként" működjenek, amelyekkel az automata szövegelemző modellek pontosságát tudjuk visszamérni. Mivel a manuális kiértékelésekben is lehet hiba, ezért a szakirodalommal összhangban az annotációkat több (két) elemző végzi és csak a megegyező értékeket vesszük figyelembe. (A humán annotációknak további szerepe van még a BERT-típusú modellek esetében, ahol a felügyelt tanítási lépésben tanulómintaként szolgálnak.) Esetünkben az annotációkat jegybanki kollégák végezték el. Felmerülhet, hogy jegybanki elemzők a piaci elemzőknél jobb minőségű annotációkat készíthetnek, de ez nem szükségszerű, jegybanki elemzők ezeknél a feladatoknál ugyanakkora információval rendelkeznek (a konkrét mondat vagy közlemény, amit értékelniük kell), mint bárki, aki ezeket a publikus tartalmakat elemzi.

[8] Elérhető a szerzők projekt weboldalán: https://gitlab.com/central_bank_tonality/kocsis_matraipitz_2025_codes. A kiegészítő adatok esetében is hangsúlyoznánk, hogy azok nem hozhatók összefüggésbe az MNB hivatalos álláspontjával.

indultunk ki saját annotációink elkészítéséhez. *Gorodnichenko et al.* (*2023*) tanulmányából az 1997 és 2010 közötti amerikai jegybank FOMC-döntések utáni közlemények szövegeit – összesen 1 243 mondatot – és az azokra publikált 3-állapotú (laza/semleges/szigorú) kiértékeléseit használtuk. *Niţoi et al.* (*2023*) négy jegybank (lengyel, magyar, román és szlovák) 2007–2022 közleményszövegeiben található mondatainak egy mintáját annotálta és publikálta. Ezek közül az MNB mondatait kiszűrtük, hogy ne legyen duplikáció a saját mintánkkal, és hogy ne növekedjen tovább az MNB-mondatok túlsúlya a többi jegybankéhoz képest, így összesen ebből a forrásból a cseh, lengyel és román jegybank kommunikációinak 1 398 mondatát használjuk. A teljes mintánk tehát 5 106 jegybanki kommunikációs mondatot tartalmaz, ennek nagyjából a felét (48,3 százalékot) teszi ki az MNB minta.[9]

Mindegyik közlemény (mondat) esetében két-két elemzőt kértünk meg az MNB-stábból, hogy a szövegeket kiértékelje aszerint, hogy a mondat a három vizsgált makrogazdasági fundamentum szempontjából melyik információtartalmat közvetíti:

– MPOL (monetáris politika):[10] (pozitív) szigorú, restriktív („héja") monetáris politikára utaló információ; (semleges) van információ, de az leginkább semleges politikára utal; (negatív) laza, expanzív („galamb") monetáris politikára utaló információ; (nincs fundamentum) nincs információ az adott témáról;

– REAL (reálgazdasági aktivitás):[11] (pozitív) kedvező reálgazdasági információ, ami erős konjunktúrára utal; (semleges) olyan reálgazdasági információ, ami leginkább semleges a konjunktúra szempontjából; (negatív) kedvezőtlen reálgazdasági információ, ami gyengébb konjunktúrára utal; (nincs fundamentum) nincs információ az adott témáról;

– INFLA (infláció):[12] (pozitív) magasabb vagy gyorsuló árdinamikára utaló információ; (semleges) stagnáló és/vagy hosszú távon átlagosnak tekinthető árdinamikára utaló információ; (negatív) alacsonyabb, vagy lassuló árdinamikára utaló információ; (nincs fundamentum) nincs információ az adott témáról.

[9] Az MNB esetében teljes közlemények kerültek az adatbázisba, így ennél a mintánál a mondatok kontextusát is pontosan ismerjük. A két tanulmányból átvett szövegek azonban szerzőik által véletlenszerűen kiemelt mondatokat tartalmaznak azok pontos forrásmegjelölése nélkül, így itt a szövegkontextus nem használható. Minden használt közleményszöveg angol nyelvű. Az MNB esetében a hivatalos angol nyelvű közleményeket használtuk, és a *Niţoi et al.* (*2023*) annotációk alapjául vett szövegek is az angol nyelvű kommunikációk.

[10] Ideértve a kamatpolitikát, a kötelező tartalékpolitikát, a devizapiaci intervenciós politikát, a jegybanki likviditásösztönző programokat, a mennyiségi lazítási programokat.

[11] Ide értjük a GDP-növekedésre, nemzetgazdasági reálkonjunktúrára, összjövedelemre, outputra, ipari termelésre, beruházásokra, ingatlanpiaci forgalomra, fogyasztásra, kiskereskedelmi forgalomra, munkanélküliségre, foglalkoztatottságra vonatkozó információkat. Az exportra és fiskális kiadásokra történő utalásokat azonban csak akkor tekintjük relevánsnak, ha explicit módon GDP-komponenseként érkezik rá utalás.

[12] Ide értjük a fogyasztói és termelői árakat, béreket (kivéve, ha reálbérekre érkezik a hivatkozás), ingatlanárakat.

A *Gorodnichenko et al.* (*2023*) és *Niţoi et al.* (*2023*) tanulmányokhoz mellékelt humán annotációkat az elemzőink figyelembe vehették saját annotációjuk elkészítéséhez, mivel azonban ezek nem voltak elérhetők az általunk meghatározott fundamentumonkénti bontásban, legalább a topik információval ki kellett egészíteniük az eredeti annotációkat. Emellett felülbírálhatták az eredeti tonalitás annotációkat is, amennyiben nem értettek velük egyet.

Végül minden mondat és fundamentum esetében (*1. táblázat*), ahol a két elemző annotációja megegyezett, az annotációt mentettük, nem egyező annotáció esetében a mondatot (az adott fundamentumra) kivettük a mintából. Az annotációk nagyjából az esetek 15–20 százalékában nem egyeztek meg, így a minta 80–85 százaléka maradt meg (5 106 mondatból 4 164, 4 175 és 4 328 mondat a három fundamentum, MPOL, REAL, INFLA esetében).

1. táblázat
Használt humán annotációk megoszlása

	Megfigyelések száma			Minta részaránya		
	MPOL	**REAL**	**INFLA**	**MPOL**	**REAL**	**INFLA**
Pozitív	251	499	471	4,9%	9,8%	9,2%
Semleges	176	171	268	3,4%	3,3%	5,2%
Negatív	400	538	344	7,8%	10,5%	6,7%
Nincs fundamentum	3 337	2 967	3 245	65,4%	58,1%	63,6%
Eltérő osztályozás (=érvénytelen)	942	931	778	18,4%	18,2%	15,2%
Összesen	5 106	5 106	5 106	100,0%	100,0%	100,0%
Ebből: Érvényes						
MNB-közlemények	1 993	2 018	2 048	47,9%	48,3%	47,3%
Fed FOMC-közlemények	886	943	1 057	21,3%	22,6%	24,4%
CNB-jegyzőkönyvek	562	537	559	13,5%	12,9%	12,9%
NBP-jegyzőkönyvek	364	338	348	8,7%	8,1%	8,0%
NBR-jegyzőkönyvek	359	339	316	8,6%	8,1%	7,3%
Összes	4 164	4 175	4 328	100,0%	100,0%	100,0%

Megjegyzés: A táblázat a tanulmányunkban használt humán annotációs minta megoszlását mutatja oszlopokban fundamentumok (MPOL: monetáris politika, REAL: reálgazdasági aktivitás, INFLA: infláció) szerint, illetve sorokban az annotált értékkategóriák és jegybankonkénti megoszlás szerint. A „pozitív"/„negatív" osztályzás az MPOL esetén szigorítást/lazítást, a REAL esetén a konjunktúra javulását/romlását, az INFLA esetén az infláció gyorsulását/lassulását jelöli. A „semleges" érték a változatlanságot vagy a historikus átlagos értékeket jelöli. A „nincs fundamentum" osztályozás jelöli, amikor az annotálók szerint a fundamentumról nem volt információ a mondatban. Az „Eltérő osztályozás" azokat a mondatokat jelöli, ahol a két annotáló osztályozása különbözött, ezeket az annotációkat érvénytelennek tekintjük, és nem használjuk a tanításban/tesztelésben.

Forrás: A szerzők számításai

2.2. Szövegelemző módszerek

2.2.1. Szabályalapú módszerek

A közgazdasági szakirodalomban és azon belül a jegybanki kommunikációval foglalkozó szövegelemző szakirodalomban 10–15 évig a szabály- vagy szótáralapú modellek voltak meghatározók. Ezek a módszerek a fundamentum topikjainak és tonalitásának azonosítását a szövegben az alapján végezték, hogy az adott szövegben megtalálható-e az adott topik/tonalitás szabályának/szótárának megfelelő kifejezések (például tipikusan az „alapkamat" és az „emelkedés" szavak az MPOL fundamentumot és a szigorítás tonalitást jelölnék ki).

Négy ilyen jellegű módszert replikáltunk és futtattunk a jegybanki kommunikációs szövegeinken.[13] *Evdokimova et al.* (*2023*) és *Mátrai-Pitz – Siket* (*2023*) módszerei egymáshoz nagyon hasonló alkalmazások, mindkettő a fundamentumok topikjaira és tonalitásukra is szólistákat tartalmaznak, mindkettő foglalkozik olyan buktatókkal, hogy a foglalkoztatottság/munkanélküliség esetén ellentétes az előjel, illetve foglalkoznak a tagadószavak dilemmáival. Amellett, hogy eltérő szótárakkal dolgoznak, a két módszer között a legfőbb különbség az, hogy míg *Evdokimova et al.* (*2023*) mondat, illetve tagmondat szintjén vizsgálódik, addig *Mátrai-Pitz – Siket* (*2023*) a mondathatárokat figyelmen kívül hagyva egy ablakos, kulcsszó-közeli elemzést végez. *Mátrai-Pitz – Siket* (*2023*) a REAL és INFLA fundamentumokra érhető el, *Evdokimova et al.* (*2023*) elérhető mindhárom fundamentumra.[14] A harmadik, szabályalapú módszer *Fulop – Kocsis* (*2023*)[15] leginkább annyiban különbözik az előző kettőtől, hogy a topikokra és tonalitás kifejezéseknél reguláris kifejezések használatával implementálja, hogy egy-egy szó helyett meghatározott távolságra lévő szavakból álló kifejezések azonosítsák a fundamentumokat, ami bonyolultabb szerkezeteket is megenged. Ez a módszer másrészt a piaci hatások lemérése céljából igyekszik a fundamentumokat kiemelni a szövegből, ezért a szabályok csak az előretekintő információkat próbálják megragadni. A negyedik módszer, *Picault – Renault* (*2017*) az EKB kommunikáció szövegei alapján generált n-gram-szótárt tartalmaz a REAL és MPOL fundamentumokra. Az előző módszerhez hasonlóan ez is hosszabb szóösszetételeket, kifejezéseket enged meg, de itt a kifejezések (csak szomszédos szavak) mintabeli előfordulása alapján automatikusan történik a kifejezés-szótár összeállítása, nem manuálisan, humán szakértők által.[16]

[13] Választásunkban olyan újabb módszereket preferáltunk, amelyek az általunk vizsgált fundamentumokra mondatszinten adtak becslést. Egy fontos referencia a szakirodalomban, *Apel – Blix-Grimaldi* (*2012*), például azért nem került be a választott módszerek közé, mert egyrészt közleményszinten és nem mondatszinten optimalizálta a keresési módszert, másrészt az általuk (és sok őket követő tanulmány által) becsült héja-galamb becslések összevontan adnak becslést REAL és INFLA fundamentumok tonalitására.

[14] *Evdokimova et al.* (*2023*) módszerében a REAL és MPOL fundamentumok is több altopikra vannak bontva, amelyeket értelemszerűen aggregálunk.

[15] A publikált verzióhoz képest az itt használt kód az INFLA-fundamentumra is tartalmaz szabályokat, amelyek a többi fundamentumhoz hasonló logikát követnek.

[16] A szerzők lexikonját használtuk, és a módszerüket Matlabban replikálva előállítottuk az „EC" (gazdasági) és „MP" (monetáris politikai) hangulatkomponenseket, amelyeket ezután a REAL és MPOL tonalitási osztályokba aggregáltunk mondatszinten az alkalmazásunk által megkövetelt kimenetnek megfelelően.

2.2.2. BERT-típusú felügyelt tanítási módszerek

A szabályalapú módszereknél lényegesen pontosabb szövegelemzési teljesítményt tudnak elérni a transzformereken alapuló nagy nyelvi modellek módszerei. Ezek a jegybanki kommunikáció területén is egyre népszerűbbek. Az egyik jellemző eljárás a doménadaptáció során a nagy nyelvi modellek közül tipikusan *Devlin et al.* (*2019*) által fejlesztett BERT-, vagy az ahhoz hasonló RoBERTa (*Liu et al. 2019*) modelleket veszik alapul. Ezeket a már előtanított alapmodelleket esetenként tovább tanítják az adott szakterület nyelvezetére. A doménadaptációhoz nagy mennyiségű, a szaknyelvre jellemző (esetünkben gazdasági/jegybanki) szöveg és nagy számítási erőforrás szükséges.

Végül a modelleket (akár a BERT-típusú alapmodellek, akár a BERT-típusú alapmodellből domén-adaptált modellek) a konkrét feladatra (pl. jegybanki szövegek értelmezése, annotációja) kell megtanítani. Ez a lépés egy felügyelt tanítási lépés, amit a modellek finomhangolásának neveznek. A modellek finomhangolásához tanítóminta szükséges, ami jellemzően humán annotációkon alapul.

A szakirodalomból négy olyan modellt választottunk, amely az aktuális jegybanki kontextushoz leginkább illeszkednek, és amely modellek mögött lévő publikációk az elmúlt években több hivatkozást kaptak. Az egyik a BIS szerzői által közzétett modell (*Gambacorta et al. 2024*), amely a RoBERTa alapmodellt kifejezetten jegybanki szövegekre domén-adaptálta (RoBERTa+Sp+Pa).[17] A szerzők a *Gorodnichenko et al.* (*2023*) által publikált Fed-kommunikáció mondatainak héja/galamb annotációját használták finomhangolásra. *Pfeifer – Marohl* (*2023*) a BERT-, RoBERTa- és más modelleket finomhangolnak jegybanki kommunikációs szövegeken, amelyek közül a finomhangolt RoBERTa-modelljük teljesített a legjobban, ezt CentralBankRoBERTa-modell néven publikálták, tanulmányunkban ez a második BERT-típusú modell, amit használunk.[18]

A harmadik és negyedik modell alapját FinBERT-modellek alkotják. Egyrészt használjuk *Huang et al.* (*2022*) FinBERT-tone modelljét, amely a BERT-architektúrát előtanítja/doménadaptálja 4,9 milliárd token pénzügyi szöveg (vállalati bejelentések, pénzügyi elemzések, eredménybeszámolók) segítségével.[19] Másrészt használjuk *Gössi et al.* (*2023*) FinBERT-FOMC modelljét, ami egy másik FinBERT-modellt (*Araci 2019* által Reuters híreken doménadaptált BERT-modell) finomhangol FOMC-jegyzőkönyvek annotációi (negatív/semleges/pozitív tonalitás) segítségével.[20]

[17] A https://bis.org/publ/work1215.htm oldalon közzétett modellek közül azt a verziót használtuk, mely során a tanulmányok mellett a beszédeket is felhasználták a doménadaptációhoz.
[18] Ez a modell tehát nem tartalmaz doménadaptációt, csak finomhangolást. Ugyanakkor ilyen csak finomhangolt modellekből is érdemes lehet kiindulni a saját finomhangolásunkhoz olyan esetekben, amikor az előzetes finomhangolás a saját alkalmazásunkhoz hasonló. Ekkor a modell még az előző szerzők tanítómintájának információit is tartalmazza. A felhasznált modell a https://huggingface.co/Moritz-Pfeifer/CentralBankRoBERTa-sentiment-classifier oldalon érhető el.
[19] A felhasznált modell a https://huggingface.co/yiyanghkust/finbert-tone oldalon érhető el.
[20] A felhasznált modell a https://huggingface.co/ZiweiChen/FinBERT-FOMC oldalon érhető el.

A négy választott modell finomhangolását 3 323 humán annotált mondat felhasználásával végeztük a fő eredményekhez (illetve a teljes MNB- és nem-MNB-minta nagysága: 2 465 és 2 641 a *3.3. fejezetbeli* mintabontás esetében), melynek véletlenszerűen kiválasztott 20 százalékát használtuk validációra a tréning során. Annotált adatbázisunknak ezt a részét nem használtuk fel a módszerek teszteléséhez.

A BERT-család modelljeinek finomhangolása során a többcímkés elrendezésnek megfelelően egyetlen encoder fölé olyan kimeneti fejet építettünk, amely a monetáris politikai, reálgazdasági és inflációs dimenziókra egymástól független döntést hoz. A többcímkés megoldás közös nyelvi reprezentációt tanít, miközben feladatonként külön döntést hoz, így jobban illeszkedik a valós tartalomszerkezethez és hatékonyabban hasznosítja az adatmennyiséget. A modellezett nyelvi jelenségek (polarizáló igék, módosítók, negáció, bizonytalansági kifejezések) ugyanis kölcsönösen segítik egymás, emiatt a modell kevesebb mintából stabilabb, általánosabb mintázatokat tanulhat. Több robusztussági elem is támogatja a finomhangolás minőségét: (i) súlyozott veszteség a ritka osztályok és a monetáris politika feladat kiemelésére, (ii) részleges címkézés kezelése, (iii) makro-F1 alapú modellkiválasztás és korai megállás a túlilleszkedés elkerülésére.[21]

2.2.3. GPT-modelleken alapuló generatív MI-módszerek

A transzformereken és nagy nyelvi modelleken alapuló módszerek másik népszerű csoportja a generatív mesterséges intelligencia használata felé fordul. A generatív mesterséges intelligencia modelljeinek – mint például az OpenAI (3.5 verzió feletti) GPT-modelleknek[22] az ígérete, hogy a felügyelt tanítási módszerekhez képest a modell mérete és előtanítása a legtöbb feladatra már általánosan használható, így elég (egy jól megírt) promptban a konkrét feladatot leírni, nem kell az előző pontban leírt felügyelt tanítási, finomhangolási lépéseket megtenni. Ezáltal megspórolható az ezzel járó humán annotáció, nem kell számítási kapacitás a finomhangolás futtatáshoz, és a modell meghívásához sem kell nagyobb gép, mert a GPT-modellek az OpenAI szerverein futtathatók. A bonyolultabb eseteknél néhány input-output példa közlését javasolják a promptban (few shot learning), illetve még bonyolultabb eseteknél a Chain-of-Thought (gondolati lánc) módszer lett népszerű (*Wei et al. 2022*), amely kisebb, egymásra épülő lépésre bontja a feladatot (tipikusan a részfeladatok magyarázatát is kéri), ami miatt a modell több időt/energiát szán

[21] A tanításban két szinten is alkalmaztunk súlyozást. Egyrészt osztályszinten kompenzáltuk az egyensúlytalanságot – különösen a szélső kategóriákat (pozitív, negatív) –, hogy a modell ne torzítson a többségi, „nincs információ" irányba. Másrészt feladatszinten a monetáris politika tévesztés nagyobb súlyt kapott a veszteségben, mivel ezt a feladatot rendszerint rosszabb teljesítménnyel sajátította el a modell. A részleges címkézés kezelésével azok a mondatok is hasznosulnak, ahol hiányzik egy vagy két dimenzió. A veszteség a feladatok keresztentrópiáinak súlyozott átlaga, amely a hiányzó címkéket kihagyja, így a részleges annotációk mellett is stabil a tanulás.

[22] Számos más generatív mesterséges intelligencia termék bontakozott ki a GPT-modellek mellett (Google Gemini, Anthropic Claude, MS Copilot, DeepSeek). A GPT választását előfizetésünk indokolja. Az ingyenes generatív MI-modellek (pl. Meta Llama) közül azok, amelyeket a számítástechnikai kapacitásaink megengednének, lényegesen gyengébb teljesítményt nyújtanak, így nem kerülnek az összevetésbe.

a megoldásra. Esetünkben a Chain-of-Thought módszer alkalmazása kézenfekvő (például a topik azonosítás az egyik lépés, tonalitás egy másik és több egyéb információtípus azonosítása szintén külön lépések, ahogy a *4. fejezetben* erre kitérünk).

Az OpenAI 2024 második felétől ún. reasoning modelleket is publikál (o1, o3, o4, majd a GPT-5 szériák), amelyek ígérete, hogy saját maguk elvégzik a Chain-of-Thought lépéseket. A tanulmányban az íráskor aktuálisan legjobb elérhető, 2025. augusztusban publikált reasoning (GPT-5) és nem-reasoning (GPT-4.1) szériák modelljeit használtuk, mindegyik esetében három-három modellméretre (nano, mini és normal) futtattuk a feladatokat. A nagyobb modellek jobb teljesítményt kínálnak, de hosszabb futtatási időt és nagyobb költségeket igényelnek.

2.3. Használt kiértékelési metrikák

A módszerek kiértékeléséhez és összehasonlításához olyan statisztikai módszereket használunk, amelyek leginkább illenek a diszkrét, többosztályos és kiegyensúlyozatlan adatainkhoz. Alkalmazásunkban minden mondat négy diszkrét értéket vehet fel (pozitív, semleges, negatív, nincs fundamentum) mindhárom fundamentum esetében, amelyek közül a „nincs fundamentum" osztály a minta meghatározó része (*1. táblázat*). Az is elvárásunk a kiértékelési metrika szempontjából, hogy a visszahívás (angolul recall, az adott címkéhez tartozó valós elemek hány százalékát sikerült megtalálni, és hány százalékát mulasztotta el – a másodfajú hiba) és precizitás (angolul precision, az adott címke összes becslése hány százalékban volt pontos és hány százalékban téves – az elsőfajú hiba) statisztikai tulajdonságokkal egyaránt foglalkozzon. Ezeknek a szempontoknak leginkább a makroátlagolt F1-statisztika felel meg (*Grandini et al. 2020*).[23]

A makroátlagolt F1-statisztika az értékkategóriánként ($i=1,\ldots,K$) vett F1-statisztikák számtani átlaga, az egyes értékkategóriákhoz tartozó F1-statisztika ($F1_i$) a kategóriánkénti bináris visszahívás ($TruePoz_i/(TruePoz_i+FalseNeg_i)$) és bináris precizitás ($TruePoz_i/(TruePoz_i+FalsePoz_i)$) harmonikus átlaga (*Takahashi et al. 2022*):

$$F1_i = 2\frac{precizitás_i + visszahívás_i}{precizitás_i^{-1} * visszahívás_i^{-1}} \tag{1}$$

$$\overline{F1}_{makro} = \sum_{i=1}^{K} F1_i \tag{2}$$

[23] A legnépszerűbb klasszifikációs statisztika a pontosság (az összes pontosan eltalált címke osztva az elemszámmal), aminek a problémája, hogy kiegyensúlyozatlan adatok esetén a nagy elemszámú kategória (esetünkben a „nincs fundamentum") fogja meghatározni a statisztikát. A kiegyensúlyozott pontosság (balanced accuracy) ezt korrigálja, ugyanakkor a visszahívásra korlátozódik, így az elsőfajú hibát figyelmen kívül hagyja. Ennek ellenére a kiegyensúlyozott pontosság statisztikát is számoljuk minden eredményünknél, azok többnyire az F1-statisztikához hasonlók.

Takahashi et al. (*2022*) kidolgozott aszimptotikus sztenderd hibákat a makroátlagolt F1-statisztikára, mi azonban a kiegyensúlyozatlan minta és egyes kategóriák esetében alacsony elemszám miatt bootstrap konfidencia-intervallumokat használunk.[24]

A módszerek kiértékelését a teljes minta véletlenszerűen választott 40 százalékos részmintáján végezzük el, ahol az MNB esetében teljes közlemények, a többi jegybank esetében pedig a mondatok 40 százalékát húzzuk.

3. Eredmények

3.1. A modellek klasszifikációs teljesítményének első értékelése

Tanulmányunk elsődleges eredményeit, a módszerek makroátlagolt F1-mutatóit a *2. táblázatban* mutatjuk be.

Az F1-metrika alapján egyértelmű hierarchia rajzolódik ki a modellcsaládok között. A legmagasabb értékeket minden fundamentum esetén a BERT-típusú modellek érik el. A (bootstrappelt 95 százalékos) konfidencia-intervallumok alapján ezek a modellek szignifikánsan jobbak valamennyi szabályalapú és a legtöbb GPT-modellnél is. A REAL és INFLA fundamentumok esetében még a leggyengébb BERT-típusú modellek F1-értékei is szignifikánsan magasabbak a GPT-család F1-értékeihez képest, leszámítva a GPT-5 modellt, amihez képest szintén magasabbak az F1 pontbecslések, de a konfidencia-intervallumok itt átfedik egymást. A GPT-típusú modellek a szabályalapú modellekhez képest szignifikánsan magasabb F1-statisztikákkal rendelkeznek mindegyik fundamentum esetében.[25]

A BERT-család jobb predikcióinak hátterében általánosságban az állhat, hogy ezek a feladatra célzottan finomhangolt modellek jobban illeszkednek a doménhez és a címkézési szabályokhoz, mint az online, általánosabb célú GPT-modellek. A finomhangolt modellek megtanulják adataink jegybank-specifikus stílusát, a mondatok annotációra legjobban illeszkedő mondatszerkezeteket (beleértve ad absurdum az annotációk szisztematikus hibáit is). Emellett jellemzőjük a magasabb fokú determinisztikusság és konzisztencia. Ezzel szemben a GPT-modellek univerzális célú, nem kifejezetten az adott feladatra optimalizált modellek, és bár a lényegesen nagyobb modellméret, a feladatspecifikus promptok és a CoT, reasoning technikák javítják az adott feladatban nyújtott teljesítményüket, így is alulmaradnak a BERT-család finomhangolt modelljeihez képest.

[24] Ehhez Jacob Gildenblat kódját vettük alapul: https://github.com/jacobgil/confidenceinterval

[25] Hangsúlyozni szeretnénk, hogy az eredeti szabályalapú módszerek outputjait azok szerzői más alkalmazásokhoz (és más adatkészletekhez) fejlesztették ki, így azokat a saját kutatásunk által megkövetelt kimeneti változókhoz kellett igazítani. Ez hátrányt jelent a BERT-típusú és GPT-alapú modellekhez képest, amelyeket kifejezetten jelen kutatás feladataira hoztuk létre, ami a szabályalapú módszerek gyengébb teljesítményéhez hozzájárulhatott.

Ezek az eredmények nagyrészt illeszkednek a szakirodalom eredményeihez. *Kim et al.* (*2024*) Fed FOMC jegyzőkönyvek mondatainak pozitív/semleges/negatív szentimentjeinek azonosításában hasonlított össze modelleket. Eredményei alapján a szabályalapú VADER-módszer F1-értékeinél érdemben magasabbak voltak a BERT-típusú modellek, és a FinBERT-FOMC modell F1-értéke náluk is megelőzte a GPT-4 eredményét, igaz, elmaradt a Llama-modellek F1-értékétől. *Huang et al.* (*2022*) szintén magasabb F1-értékeket mutat a BERT-típusú modellek esetében más szabályalapú modellekhez képest. *Gambacorta et al.* (*2024*) a monetáris politikai tonalitás azonosításban bemutatja, hogy az általuk használt RoBERTA+Sp+Pa modell pontossága a legtöbb esetben felülmúlta az akkori sztenderd GPT (3.5 és 4 széria) és más generatív nagy nyelvi modellek (Llama, Mistral) teljesítményét.[26] *Pfeifer – Marohl* (*2023*) eredményei alapján szentiment-klasszifikációban a BERT-típusú modellek jobban teljesítettek a szabályalapú gépi tanulási módszerekkel összevetve.

Modellcsaládok között általában nagyobbak a különbségek, mint modellcsaládokon belül. A BERT-típusú modellek F1-konfidenciaintervallumai alapján a különbségek nem érik el a szokásos statisztikai hibahatárt, de pontbecslések alapján és a három fundamentum összességében *Huang et al.* (*2022*) FinBERT-tone modellje tűnik a legjobbnak. Az infláció esetében az F1-érték *Pfeifer – Marohl* (*2023*) CentralBank-RoBERTa modellje esetében a legmagasabb.

A GPT-modellek közötti összehasonlításokban kevés meglepetés van. Az egyre nagyobb modellméret javítja az F1-statisztikákat (a nano- és minimodellek között lényegesen nagyobb a különbség, mint a mini- és normál modellek között). A Chain-of-Thought modellek gyengébbek a Reasoning-modelleknél. A GPT-modellek közül így a GPT-5 modell produkálja a legmagasabb F1-statisztikát mindhárom fundamentum esetében.

A szabályalapú modellek közül az MNB szerzőinek (*Mátrai-Pitz – Siket 2023*) a módszere emelkedik ki a többi közül mindkét fundamentumnál, amelyikre ez a módszer elérhető (REAL, INFLA). A reguláris kifejezések alapú modell (*Fulop – Kocsis 2023*) és *Evdokimova et al.* (*2023*) alacsonyabb, de *Mátrai-Pitz – Siket* (*2023*) módszeréhez hasonló statisztikát ér el ezeken a fundamentumokon. Az MPOL-fundamentumon valamennyi módszer gyengébb F-értékeket produkál, ami arra utalhat, hogy ennek a fundamentumnak a kifejezéstára lehet a legbonyolultabb, erre lehet legnehezebb szabályalapú azonosítást rögzíteni. *Picault – Renault* (*2017*) módszer gyengébb teljesítményét leginkább az indokolhatja, hogy ez a módszer jellegénél fogva kifejezetten egy adott mintára illeszkedik, és az általuk használt EKB kommunikáció szövegei nincsenek a mintánkban.

[26] A teljesítménykülönbség speciális esetekben megfordult: egyrészt az GPT-, Llama-modellek finomhangolásakor, másrészt, amikor mondatok helyett hosszabb hírek szövegeinek tonalitásértékelése volt a feladat, ráadásul szűkebb tanulómintán.

2. táblázat
Makroátlagolt F1 statisztikák

Szerzők	Modell jellege	MPOL	REAL	INFLA
Evdokimova et al. (*2023*)	szabályalapú	0,357	0,532	0,540
		[0,334 – 0,383]	*[0,511 – 0,551]*	*[0,518 – 0,563]*
Mátrai-Pitz – Siket (*2023*)	szabályalapú	na	0,541	0,576
		na	*[0.514, 0.574]*	*[0.550, 0.603]*
Fulop – Kocsis (*2023*)	szabályalapú, regex	0,261	0,532	0,555
		[0,245 – 0,287]	*[0,502 – 0,570]*	*[0,527 – 0,586]*
Picault – Renault (*2017*)	szabályalapú, n-gram	0,204	0,285	na
		[0,189 – 0,227]	*[0,268 – 0,303]*	*na*
Huang et al. (*2022*)	BERT-típus, FinBERT-tone	**0,794**	**0,844**	0,838
		[0,761 – 0,826]	***[0,814 – 0,874]***	*[0,811 – 0,863]*
Gössi et al. (*2023*)	BERT-típus, FinBERT-FOMC	0,763	0,823	0,809
		[0,727 – 0,795]	*[0,791 – 0,853]*	*[0,780 – 0,838]*
Pfeifer – Marohl (*2023*)	BERT-típus, C.B.RoBERTa	0,778	0,827	**0,839**
		[0,744 – 0,812]	*[0,795 – 0,857]*	***[0,812 – 0,865]***
Gambacorta et al. (*2024*)	BERT-típus, CB-LM RoBERTa	0,754	0,830	0,823
		[0,718 – 0,788]	*[0,798 – 0,860]*	*[0,795 – 0,851]*
OpenAI (2025)	CoT, GPT-4.1-nano	0,516	0,604	0,661
		[0,482 – 0,550]	*[0,577 – 0,633]*	*[0,631 – 0,692]*
OpenAI (2025)	Reasoning, GPT-5-nano	0,584	0,646	0,610
		[0,549 – 0,623]	*[0,615 – 0,686]*	*[0,582 – 0,644]*
OpenAI (2025)	CoT, GPT-4.1-mini	0,712	0,733	0,701
		[0,682 – 0,744]	*[0,700 – 0,769]*	*[0,669 – 0,733]*
OpenAI (2025)	Reasoning, GPT-5-mini	0,716	0,748	0,747
		[0,684 – 0,751]	*[0,717 – 0,780]*	*[0,716 – 0,777]*
OpenAI (2025)	CoT, GPT-4.1	0,711	0,714	0,684
		[0,677 – 0,748]	*[0,681 – 0,750]*	*[0,652 – 0,719]*
OpenAI (2025)	Reasoning, GPT-5	0,738	0,778	0,762
		[0,704 – 0,773]	*[0,746 – 0,813]*	*[0,730 – 0,792]*

Megjegyzés: A táblázat a modellbecslések humán annotációkkal szembeni kiértékelésére makroátlagolt F1-statisztikákat és az azok körüli bootstrap konfidencia-intervallumokat mutatja be 14 modell és három fundamentum esetében. Öt jegybank közleményeinek/mondatainak véletlen választott részmintáját, összesen 2 384 mondatát használtuk tesztmintának. Fundamentumonként kiemeltük a legjobb F1-statisztikát felmutató modelleket.

Forrás: A szerzők számításai

3.2. A klasszifikációs teljesítmény komponenseinek értékelése

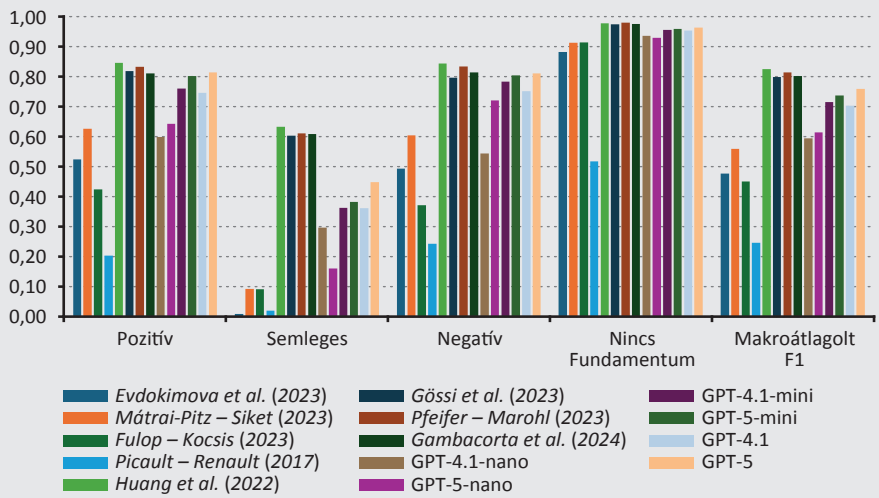
A következőkben az F1-statisztikák közötti különbségek hátterét a metrika komponensei segítségével vizsgáljuk. Legrészletesebben a *Függelék* táblázataiban közöljük a fundamentumokra, értékkategóriákra és találati/hibatípus szerint bontott megfigyelések számait.

Az *1. ábra* az aggregált F1 mutató hátterét értékkategóriákra választva (de fundamentumok között átlagolva) elemzi. Az ábra alapján valamennyi módszer legkönnyebben a „nincs fundamentuminformáció" kategóriát tudja azonosítani. A fundamentum meglétének azonosítása topikazonosítási feladat, intuitívnek tűnik, hogy ez könnyebb feladat, mint a tonalitás címkézése. Ezzel szemben valamennyi módszer számára a legnehezebb feladatnak a „semleges" kategória eltalálása tűnik. Ez is érthető, egyrészt mivel a „semleges" kategória a két másik tonalitás kategória között található (a pozitív és negatív címkék felcserélése nehezebb, mint a semleges felcserélése bármelyik másikra), másrészt a „semleges" kategória keverhető össze leginkább azokkal az esetekkel is, ahol a fundamentumra vonatkozó információ is bizonytalan.

A modellcsaládok közötti relatív különbségek jelentősek. A BERT-típusú modellek továbbra is mindegyik értékkategóriában a legjobb teljesítményt mutatják fel, de különösen nagy az előnyük a semleges kategória felismerésében a többi modellcsaládhoz képest. A szabályalapú modellek a legkisebb elmaradást a többi módszerhez képest a „nincs fundamentum" kategóriában érik el, a legnagyobb elmaradást pedig a „semleges" kategóriában.[27]

[27] A *Függelékben* közölt táblázatok alapján a „semleges" kategória elemszáma mintegy fele, harmada a pozitív és negatív kategóriáknak, de a találati arányok sokkal alacsonyabbak valamennyi módszernél. Különösen a szabályalapú módszereknél nagy a másodfajú hiba (ez a hamis negatív eset, tehát amikor a valós semleges értéket a módszer máshova kategorizálja), a 70–100 körüli semleges esetből 10 alatti találat érkezik a szabályalapú módszereknél, miközben a BERT-típusú módszereknél bőven 50 százalék feletti a találati arány, GPT-modelleknél 20–30 százalék jellemző. Ezzel szemben a „pozitív" és „negatív" kategóriáknál a szabályalapú (igaz pozitív) találati arányok 30–50 százalék körüliek, míg a BERT- és fejlett GPT-módszereknél a 70–80 százalék jellemző, a két család között kevés eltéréssel. Ez alapján leginkább a „semleges" kategórián való teljesítés határozza meg az F1-statisztikák modellcsaládok közötti különbségeket.

1. ábra
F1-statisztika kategóriák szerinti bontásban



Megjegyzés: Az ábra az F1-mutató értékeinek pontbecslését mutatja módszerenként és címkekategóriák (pozitív/semleges/negatív/nem fundamentum) szerinti bontásban a három fundamentum átlagában. Az értékek összerendezése fundamentumok között lehetséges, mert a pozitív/semleges/negatív értékek mindhárom fundamentum esetén leginkább a szigorú/semleges/laza monetáris politikai irányultsággal konzisztens.

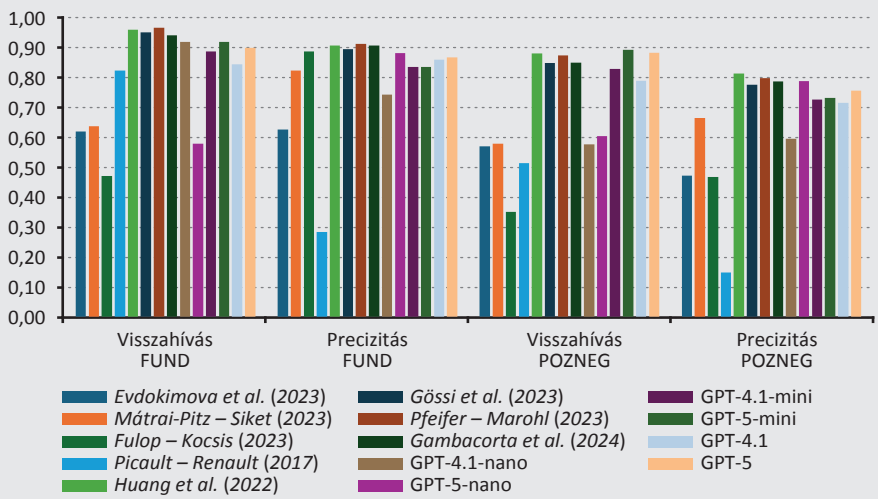
Forrás: A szerzők számításai

A makroátlagolt F1-statisztika (2) a makroátlagolt visszahívás és precizitás metrikák harmonikus átlaga, amelyek pedig az egyes értékkategóriánkénti visszahívás, illetve precizitás metrikák súlyozatlan átlagai. A *2. ábra* a visszahívás és precizitás mutatóit veti össze módszerek között, egyrészt a fundamentum azonosítása („nincs fundamentum információ" kategória fundamentumok közötti átlagolása), másrészt az előjeles tonalitások azonosítása („pozitív" és „negatív" kategóriák fundamentumok közötti átlagolása) tekintetében.

Az ábra alapján többnyire fennmarad a korábban látott sorrend a modellcsaládok teljesítményei között, ugyanakkor vannak érdekes eltérések az aggregált statisztikákhoz képest. Egyrészt a fundamentumok azonosítása tekintetében a szabályalapú modellek többnyire precizitásban sokkal jobb eredményt érnek el, mint visszahívásban, sőt *Mátrai-Pitz – Siket* (*2023*) utol is éri, *Fulop – Kocsis* (*2023*) meg is előzi ebben tekintetben a GPT-modelleket. Ezek szerint a szabályalapú modellek inkább elsőfajú hibák tekintetében gyengék (sokszor nem azonosítják a humán annotálók szerint a mondatban meglévő fundamentumokat), másodfajú hibát viszont ritkábban vétenek (a mondatokban azonosított fundamentum csak ritkán téves). Kivétel e tekintetben *Picault – Renault* (*2017*), amely pont fordítva, a visszahívásban erős.

A szabályalapú modellek a pozitív/negatív tonalitás esetében viszont általánosan gyengébbek a másik két modellcsaládnál, ami összhangban van az *1. ábrán* látott eredményekkel. A tonalitást lényegesen nehezebb megragadni mindegyik módszer számára, de még sokkal nehezebb fix szótárak alapján, mert jellemzően több szó együttese pontosítja ezt a tartalmat, és az alkotóelemeket jelentő szavak is sok szinonimával rendelkeznek, így nehéz ezeket felsorolni. A nagy nyelvi modellek viszont a szavak kontextusát is jobban ragadják meg. A GPT-modellek általában erősek ezeknek az információelemeknek a kiemelésében, de a BERT-típusú modellek a finomhangoláson (és doménadaptáción) keresztül még pontosabban tudnak ráilleszkedni az adott szaknyelvi környezetre, szakzsargonra. A GPT-modellek egyedül a tonalitás visszahívás metrikájában tudják a nagyobb reasoning modellek esetében megelőzni a BERT-modelleket (tehát a valós pozitív/negatív tonalitás nagyobb részét tudják elérni, de a precizitás gyengébb, tehát a pozitív/negatív becsült tonalitás többször bizonyul pontatlannak).

2. ábra
F1-statisztika kategóriák szerinti bontásban



Megjegyzés: Az ábra az F1-mutató értékeinek pontbecslését mutatja módszerenként és címkekategóriák (pozitív/semleges/negatív/nem fundamentum) szerinti bontásban a három fundamentum átlagában. Az értékek összerendezése fundamentumok között lehetséges, mert a pozitív/semleges/negatív értékek mindhárom fundamentum esetén leginkább a szigorú/semleges/laza monetáris politikai irányultsággal konzisztens.

Forrás: A szerzők számításai

Ami érdekesség még, hogy a mindkét ábrán bemutatott komponensek tekintetében megmarad a két FinBERT- és két RoBERTa-típusú modellek között a (pontbecslések) sorrendje (a FinBERT-tone mutatói magasabbak, mint a FinBERT-FOMC, és a CentralBankRoBERTa mutatói magasabbak, mint a RoBERTa+Sp+Pa modellé), bár a különbségek statisztikailag nem szignifikánsak.

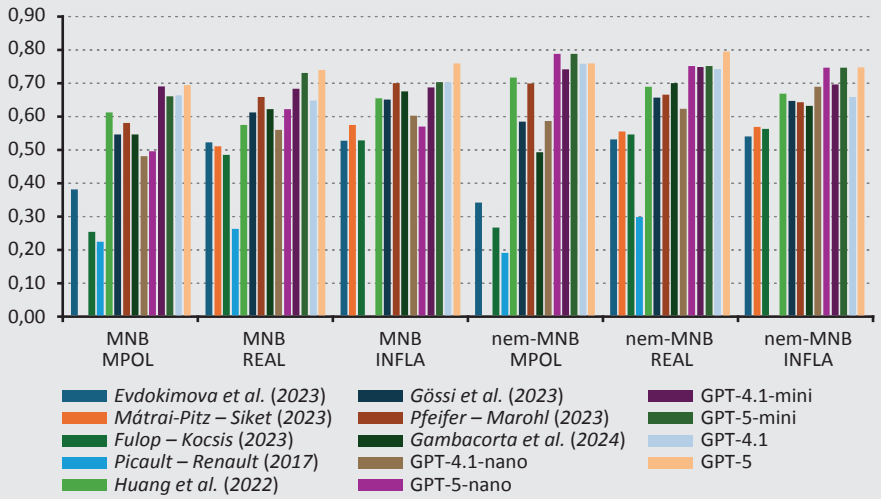
3.3. Az általánosítási képesség vizsgálata jegybankok szerinti mintamegosztás segítségével

Az eddigiekben a tesztmintát a teljes halmazból választottuk ki véletlenszerűen, így ezek az eredmények arra engednek következtetni, hogy új jegybanki kommunikáción milyen pontosságot érhetnek el a különböző módszerek. A következőkben különböző jegybanki mintákra szegmentáljuk a mintát, ami által viszont arra következtethetünk, hogy jegybankok között mennyire átvihetőek/konzisztensek az eredmények.

A mintát két részre bontottuk, egy csak MNB-közleményeket tartalmazó részmintára és a négy másik jegybank kommunikációit tartalmazó részmintára.[28] A BERT-típusú modelleket ezúttal az MNB tesztmintáin való tesztelés előtt a nem-MNB szövegeken, a nem-MNB tesztmintán való teszteléshez az MNB-szövegeken tanítottuk/finomhangoltuk. A részmintás tesztelés indikációt adhat arra, hogy a finomhangolt modellek esetében a betanítás egy adott jegybanki szövegmintán milyen klasszifikációs eredménnyel jár más jegybankok szövegei tekintetében. Hasonlóképp más módszerek esetében is ez a teszt rámutathat, ha a szabályok/promptok eltérő jegybankok között eltérő klasszifikációs eredményekhez vezet. A mintafelosztás még annyiban eredményez további (az előzőektől külön nem kezelhető) tesztet, hogy míg a nem MNB jegybanki kommunikációs szövegek független mondatok véletlenszerű mintái, az MNB-közleményeknél a mintáink teljes közleményeket együtt kezelnek, ami néhány módszer (*Fulop – Kocsis 2023*, GPT-modellek) esetében jelenthet előnyt (utóbbiak a teljes közleményt figyelembe véve annotálják mondataikat), illetve a humán annotálók is az MNB-mondatok esetén kontextusukat is figyelembe tudták venni a mondatok címkézésénél. Az MNB-közlemények a teljes minta csaknem felét teszik ki, így kellően nagy elemszám marad a tesztelésre még az alacsony részarányú értékkategóriákban is.

[28] A négy másik jegybank szétbontásakor nem lenne elég mintánk az alacsony elemszámú kategóriákban.

3. ábra
F1-statisztika részminta és fundamentumok szerint



Megjegyzés: Az ábra az F1-mutató értékeinek pontbecslését mutatja módszerenként és a három fundamentum a tesztmintát két részmintára szűkítve; csak MNB-közleményeket tartalmazó részmintára és a négy másik jegybank kommunikációit tartalmazó részmintára. A BERT-típusú modelleket ezúttal az MNB tesztmintáin való tesztelés előtt a nem-MNB-szövegeken, a nem-MNB-tesztmintán való teszteléshez az MNB-szövegeken tanítottuk/finomhangoltuk.

Forrás: A szerzők számításai

A *3. ábra* fontos tanulsága, hogy a BERT-típusú modellek klasszifikációs teljesítménye érdemben csökken, ha a tanulóminta és a tesztminta eltérő jegybankoktól származik. A BERT-típusú modellek továbbra is lényegesen erősebb klasszifikációs teljesítményt nyújtanak a szabályalapú modellekhez képest, de elmaradnak a legjobb GPT-modellek teljesítményéhez viszonyítva. Annak érdekében tehát, hogy a BERT-típusú modellek teljesítményelőnye érvényesüljön, elengedhetetlen a konkrét szövegtípusokra és feladatokra való finomhangolásuk, mivel általánosításra és más típusú adatokon való hatékony teljesítményre kevésbé alkalmasak.

Többnyire fennmarad a korábban látott modellcsaládon belüli sorrend, a FinBERT-típusú modellek közül *Huang et al.* (*2022*) F-mutatói magasabbak, a RoBERTa-típusú modellek közül pedig *Pfeifer – Marohl* (*2023*) modellje teljesít jobban. A szabályalapú modellek közül legtöbbször *Mátrai-Pitz – Siket* (*2023*) és az MPOL-fundamentum esetén *Evdokimova et al.* (*2023*) a legerősebbek, a GPT szériái közül pedig jellemzően a GPT-5 modell produkálja a legmagasabb F1-értékeket. Előzetes várakozásunk, hogy a GPT-modellek (és a *Fulop – Kocsis 2023* szabályalapú módszer) az MNB-szövegeken jobb klasszifikációs eredményeket érnek el a mondatok kontextusának

ismerete miatt, nem igazolódik, sőt az MPOL- és REAL-fundamentumok esetében az F1-értékek a nem-MNB-mintán magasabbak, míg az INFLA-fundamentumnál az eredmények vegyesek.

A fundamentumok közötti összehasonlításokban a szabályalapú módszerek esetében mindkét részmintán az MPOL F1-értékek jóval elmaradnak a másik két fundamentumon elért értékekhez képest. Ez a különbség azonban a BERT-típusú és GPT-modellek esetében nem nyilvánvaló. Az MNB-részmintán a modellek többsége ugyan valamelyest gyengébb teljesítményt ér el az MPOL-fundamentumon, de a nem-MNB-részmintán inkább a modellek vegyesen teljesítenek fundamentumok között.

3.4. Gyorsaság és költségek

A modellek választásánál további szempont lehet a futtatási sebesség, illetve a modellek futtatásának költségei. Saját tapasztalatainkat a *3. táblázat* mutatja be. Ki kell emelni, hogy a futtatási sebesség nagymértékben függ a használt hardvertől a szabályalapú módszerek és az on-prem használt BERT-típusú modelleknél, ugyanakkor nem számít a GPT-modelleknél (amelyek a felhőben futnak). A táblázatban közölt futtatási idő természetesen lényegesen lerövidíthető, ha van lehetőség a párhuzamosításra, de ahol mi alkalmaztunk párhuzamosítást, ott a futtatási szálak számával felszoroztuk a tapasztalt futtatási időt, hogy a táblázatban közölt értékek párhuzamosítás nélkül legyenek érthetők/összehasonlíthatók. A futtatási idő természetesen nem tartalmazza a modellek létrehozásának (BERT-típusúak esetében az annotáció és modelltanítás, szabályalapú és GPT-modelleknél a szabályok és promptok létrehozásának) idejét.

A költségek esetében kiemelendő, hogy a táblázatban kizárólag a GPT-futtatások költségeit mértük le, nem foglalkoztunk a főleg a BERT-típusú megoldásokhoz köthető közvetett, potenciálisan jelentős költségekkel (a nagyobb szerverkapacitásból kapcsolódó költségekkel: hardver beszerzésével, installálással és fenntartással kapcsolatos munkaerő- és energiaköltségek, humán annotáció költségei).

3. táblázat
Futtatási idő és a futtatás közvetlen költségei

	Futtatási időigény (perc)		Ár
	MNB-tesztminta	nem-MNB-tesztminta	USD
Evdokimova et al. (*2023*)	0,4	0,4	0,0
Mátrai-Pitz – Siket (*2023*)	0,3	0,1	0,0
Fulop – Kocsis (*2023*)	0,8	0,8	0,0
Picault – Renault (*2017*)	0,9	0,9	0,0
Huang et al. (*2022*)	20,3	18,4	0,0
Gössi et al. (*2023*)	16,5	16,1	0,0
Pfeifer – Marohl (*2023*)	21,6	15,0	0,0
Gambacorta et al. (*2024*)	17,9	20,2	0,0
GPT-4.1-nano	182,8	241,0	2,7
GPT-5-nano	101,0	605,1	3,1
GPT-4.1-mini	211,7	275,7	12,2
GPT-5-mini	129,5	766,7	7,4
GPT-4.1	240,9	278,3	51,8
GPT-5	209,9	1140,7	46,8

Megjegyzés: A táblázat az egyes módszerek esetében a tesztminta két részének (összesen mintegy 2 100 mondat) futtatási idejét és közvetlen futtatási költségét (USD) mutatják szerverünkön (a párhuzamosításból adódó időmegtakarítással visszakorrigáltunk, így az időigény azt mutatja, hogy szerverünkön 1 futtatási szál esetén hány perc alatt futott volna végig a tesztminta).

Forrás: A szerzők számításai

Ezeket a disclaimereket figyelembe véve megállapíthatjuk, hogy közvetlen futtatási sebesség szempontjából a szabályalapú módszerek a BERT-típusú modelleknél 10–50-szer voltak gyorsabbak. A BERT-típusú modellek sebessége között nagyságrendi különbségek nem voltak (modellcsaládon belül), hozzájuk képest a CoT GPT modellek mintegy 10-szer lassabbak voltak, a GPT reasoning modellek (GPT 5 széria) sebessége pedig nagyban függött a tesztmintától: az MNB-mintán (ahol a modell a közlemények összes mondatát egyszerre értékelhette ki) a CoT GPT modelleknél valamelyest gyorsabb, a mondatonként értékelt nem-MNB-tesztmintán viszont a CoT-modelleknél 3-4-szer lassabb volt. A futtatás közvetlen költségeiben a GPT-modellek között modellméret szerint jelentősebb különbségek voltak, kategóriánként 5-ös szorzóval lehetett számolni.

4. Kiterjesztések

A tanulmány végén szeretnénk röviden kitérni olyan kutatási irányokra, amelyeket a szakirodalomban, illetve saját tapasztalataink alapján előremutatónak látunk.

Az egyik ilyen irány, amelyet néhány tanulmányban láttunk, és amelyet saját kódjainkban is alkalmazunk, az a jegybanki kommunikáció információtartalmának több

aspektusra kiterjedő vizsgálata. A szakirodalmi sztenderd vizsgálódás továbbra is az egyszerű pozitív/negatív vagy héja/galamb skálán azonosított tonalitás. Természetesen ennél sokkal többrétegű a kommunikációban lévő információ és annak pontosabb megértése segítheti a piaci hatások visszamérését, magyarázatát és egyéb jegybanki alkalmazásokat.

Yao et al. (*2025*) a Fedspeak kommunikációban olyan makrogazdasági kauzális láncokat próbál azonosítani, amelyek a monetáris politikai irányultság mögött érthetők. *Geiger et al.* (*2025*) Chain-of-Thought prompt technikát alkalmaznak a tonalitás előjele után intenzitásának megállapítására, majd ezt a szövegben lévő monetáris politikai irányultságot meglévő egyéb fundamentumokra vonatkozó információkkal együtt (pl. a stáb inflációs előrejelzések tükrében) értékelik.

Saját GPT-promptjaink a következő információrétegek irányában lépnek tovább:

– Három fundamentumra külön azonosítjuk a tonalitást, de a fundamentumokon belül is azonosítunk topikokat. Ilyet máshol is látunk, pl. *Evdokimova et al.* (*2023*) az MPOL-fundamentum esetében azonosítja az előretekintő iránymutatás és mennyiségi lazítás komponenseket, vagy az aktivitáson belül konkretizálja a munkaerőpiaci aktivitást. Mi ennél tovább megyünk, a REAL-fundamentum esetén 7, MPOL esetén 6, INFLA esetén 4 topikot jelölünk ki.

– Megállapítjuk, hogy milyen földrajzi egységhez kapcsolódik a fundamentum-információ (nem mindegy, hogy egy jegybank a saját vagy a globális növekedési kilátásokról beszél).

– Kiemeljük a fundamentumok dinamikáját (sokszor az zavarja össze a humán és gépi annotációkat, hogy ellentmondó a szintre és változásra vonatkozó információ: pl. „a munkanélküliség kissé emelkedett, de továbbra is historikusan alacsony szinten marad”).

– Időbeliségre vonatkozó információ azonosításával próbálkozunk (múltra/jelenre/jövőre vonatkozik a fundamentum tonalitása). Ebben a tekintetben fontos kontribúció *Byrne et al.* (*2023*), amely részletesen dokumentálja, hogy az időbeliségre vonatkozó információk érdemben magyarázzák a hozamokban látott meglepetéseket.

– Tény/meglepetés/várakozások és várakozásokon belül felfelé/lefelé mutató kockázatok elkülönítése.

– Végül a tonalitásnál nemcsak előjeleket, de különböző intenzitást is kérünk a promptban (–3 és +3 közötti skála), ahogyan ezt sokan mások is megteszik.

A *4. táblázat* bemutatja *Gorodnichenko et al.* (*2023*) mintájában lévő FOMC három mondat esetében a GPT-promptjaink részletesebb outputját a GPT-5 modell által.

4. táblázat
A GPT-prompt részletesebb outputjai kiválasztott mondatokra

Mondat	Topik kód	Topik szöveg	GEO	Tone szöveg	Tone pontsz.	Tény v. várt	Dinamika	Idősík
The Committee will maintain the target range for the federal funds rate at 0 to 1/4 percent and continues to anticipate that economic conditions are likely to warrant exceptionally low levels of the federal funds rate for an extended period.	MPOL_RATE	federal funds rate	United States	range unchanged	0	OBSERV	CHANGE	CURRENT
	MPOL_RATE	federal funds rate	United States	0 to 1/4 percent	–3	OBSERV	LEVEL	CURRENT
	MPOL_FG	forward guidance	United States	low rates extended period	–2	EXPECT	LEVEL	FUTURE
	REAL_CYCLE	economic conditions	United States	weak economic conditions	–2	EXPECT	LEVEL	FUTURE
Although inflation pressures seem likely to moderate over time, the high level of resource utilization has the potential to sustain those pressures.	REAL_CYCLE	resource utilization	current country	high level of resource utilization	2	OBSERV	LEVEL	CURRENT
	INFLA_CPI	inflation pressures	current country	likely to moderate	–2	EXPECT	CHANGE	FUTURE
	INFLA_CPI	inflation pressures	current country	sustain those pressures	2	EXPECT_POSRISK	LEVEL	FUTURE
Although economic activity is likely to remain weak for a time, the Committee continues to anticipate that policy actions to stabilize financial markets and institutions, fiscal and monetary stimulus, and market forces will contribute to a gradual resumption of sustainable economic growth in a context of price stability.	MPOL_GEN	monetary stimulus	current country	monetary stimulus	–2	OBSERV	LEVEL	CURRENT
	REAL_GDP	economic activity	current country	likely to remain weak	–2	EXPECT	LEVEL	MedFUTURE
	REAL_GDP	economic growth	current country	gradual resumption	2	EXPECT	CHANGE	MedFUTURE
	REAL_GDP	economic growth	current country	sustainable	2	EXPECT	LEVEL	MedFUTURE
	INFLA_CPI	price stability	current country	price stability	0	EXPECT	LEVEL	FUTURE

Megjegyzés: A táblázat a GPT-5 modell részletesebb outputjainak (topik, földrajzi lokáció, tonalitás, tény/várakozás, dinamika: szint/változás, idősík) eredményét mutatja be három kiválasztott Fed FOMCmondat példáján.

Forrás: A szerzők számításai

Hasonlóan sokrétű információ kinyerése sem szabályalapú (túl bonyolult lenne a szabályrendszer), sem BERT-típusú módszerekkel (hatalmas humán annotált mintát igényelne, hogy a sok lehetséges értékkategória mindegyikére megfelelő elemszám maradjon) nem célravezető, így ezen a vonalon a promptokkal dolgozó generatív modelleké lehet a jövő.

Egy másik irány, amit ígéretesnek tartunk és felmerül a szakirodalomban is (pl. *Geiger et al. 2025*), hogy a BERT-típusú modellek finomhangolásához költséges humán annotációk helyett/mellett generatív nagy nyelvi modellek annotációit használjuk. Korábbi eredményeink alapján a BERT-típusú modellek akkor tudnak igazán pontos klasszifikációs teljesítményt nyújtani, ha az adott jegybank szövegeire vannak specifikusan finomhangolva. Mivel a humán annotációk beszerzése időigényes és költséges lehet, esetenként egyszerűbb lehet a címkézést valamely erősebb online generatív modellel (pl. OpenAI nagy GPT-modelljeivel) elvégezni. A néhány száz/ezer mondat alapján betanított BERT-típusú modellek ezután a GPT-modelleknél már gyorsabban és előfizetési költségek nélkül futtathatók a szélesebb szövegadatbázis kiértékelésére.

5. Konklúziók

Tanulmányunk a jegybanki szövegelemzés egyre szélesebb szakirodalmához kíván hozzájárulni azáltal, hogy összeveti három különböző modellcsalád (szabályalapú, BERT-típusú és GPT-modellek) különböző tulajdonságait három makrogazdasági fundamentum tekintetében, öt jegybank szövegeit felhasználva. Az eredmények alapján, bár klasszifikációs teljesítményükben lényeges különbségek vannak, mindhárom modellcsalád használatának lehet létjogosultsága az adott szövegelemzési feladat humán-, IT-, pénzügyi- lehetőségei és különböző időkorlátok mellett.

A szabályalapú módszerek lényegesen gyengébb klasszifikációs teljesítményét esetenként kompenzálhatja a módszerek gyorsasága (amely egy nagyságrenddel jobb a BERT-típusú és két nagyságrenddel a GPT-modelleknél), valamint a módszer alacsony költsége (nem kellenek humán annotációk, kis hardver-igény) és átláthatósága (nincs feketedoboz jelleg).

A szakirodalom eredményeivel összhangban a BERT-típusú modellek megfelelő humán annotáció mellett a legerősebb GPT-modelleknél is jobb teljesítményt érnek el és azoknál lényegesen gyorsabbak, offline futtathatók, így előfizetést sem igényelnek. Ezeknek a modelleknek a kiemelkedő teljesítményéhez fontos az adott jegybanki szöveghez igazított humán annotált tanítóminta (más jegybanki szövegeken már elveszíti a GPT-vel szembeni előnyét), és a futtatáshoz nagyobb hardver szükséges, mint a másik két modellcsaládhoz.

A GPT-modellek mérettől és szériától függően teljesítenek, az újabb, nagyobb modellek a jobb klasszifikációs teljesítményt magasabb futtatási költségek és hosszabb futási idő mellett biztosítják. A GPT-modellek lényeges előnye, hogy nincs szükség humán annotációkra, a promptok főleg az újabb reasoning modellek esetében rugalmasan és a szöveg sokféle információ aspektusának a kinyerésére alkalmazhatók. Továbbá, amennyiben egy újabb, jobb teljesítményű modell jelenik meg, az különösebb fejlesztési igény nélkül beépíthető meglévő folyamatainkba.

Bizonyos esetekben célszerű lehet modellcsaládok kombinációit használni. Például a drágább és lassabb nagy GPT-modelleket érdemes lehet kisebb annotációs minták elkészítésére felhasználni, amelyen BERT-típusú modelleket tanítva később nagyobb adatbázisokat lehet gyorsabban és költséghatékonyan elemezni.

Felhasznált irodalom

Apel, M. – Blix Grimaldi, M. (2012): *The Information Content of Central Bank Minutes*. Working Paper No. 261, Sveriges Riksbank. https://doi.org/10.2139/ssrn.2092575

Araci, D. (2019): *FinBERT: Financial Sentiment Analysis with Pre-Trained Language Models*. arXiv preprint arXiv:1908.10063. https://doi.org/10.48550/arXiv.1908.10063

Bihari Péter (2015): *Odüsszeuszi utazás – az előretekintő iránymutatás tapasztalatai*. Közgazdasági Szemle, 62(4): 749–766. https://www.kszemle.hu/tartalom/cikk.php?id=1568

Bihari Péter – Sztanó Gábor (2015): *Sikertörténet vagy sok hűhó semmiért?.* Köz-Gazdaság, 10(2): 23–39. https://unipub.lib.uni-corvinus.hu/2033/

Blinder, A.S. – Ehrmann, M. – Fratzscher, M. – De Haan, J. – Jansen, D.J. (2008): *Central Bank Communication and Monetary Policy: A Survey of Theory and Evidence*. Journal of Economic Literature, 46(4): 910–945. https://doi.org/10.1257/jel.46.4.910

Brown, T.B. – Mann, B. – Ryder, N. – Subbiah, M. – Kaplan, J. – Dhariwal, P. et al. (2020): *Language Models are Few-Shot Learners*. Advances in Neural Information Processing Systems, 33: 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

Byrne, D. – Goodhead, R. – McMahon, M. – Parle, C. (2023): *The Central Bank Crystal Ball: Temporal Information in Monetary Policy Communication.* Research Technical Paper Vol. 2023 No. 1, Central Bank of Ireland. https://ideas.repec.org/p/cbi/wpaper/1-rt-23.html

Correa, R. – Garud, K. – Londono, J.M. – Mislang, N. (2017): *Sentiment in Central Banks" Financial Stability Reports.* International Finance Discussion Paper No. 1203. https://doi.org/10.17016/IFDP.2017.1203

Csortos Orsolya – Lehmann Kristóf – Szalai Zoltán (2014): *Az előretekintő iránymutatás elméleti megfontolásai és gyakorlati tapasztalatai.* MNB-Szemle, 9(2): 45–55. https://www.mnb.hu/letoltes/csortos-orsolya-lehmann-kristof-szalai-zoltan.pdf

Devlin, J. – Chang, M.W. – Lee, K. – Toutanova, K. (2019): *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. In: Burstein, J. – Doran, C. - Solorio, T. (eds.): Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, June, pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423

Dovern, J. – Fritsche, U. – Slacalek, J. (2012): *Disagreement among Forecasters in G7 Countries*. Review of Economics and Statistics, 94(4): 1081–1096. https://doi.org/10.1162/REST_a_00207

Ehrmann, M. – Eijffinger, S. – Fratzscher, M. (2010): *The role of central bank transparency for guiding private sector forecasts.* ECB Working Paper No. 1146. http://dx.doi.org/10.2139/ssrn.1532312

Eusepi, S. – Preston, B. (2010): *Central Bank Communication and Expectations Stabilization*. American Economic Journal: Macroeconomics, 2(3): 235–271. https://doi.org/10.1257/mac.2.3.235

Evdokimova, T. – Nagy-Mohacsi, P. – Ponomarenko, O. – Ribakova, E. (2023): *Central Banks and Policy Communication: How Emerging Markets have Outperformed the Fed and ECB*. Peterson Institute for International Economics Working Paper No. 23–10. https://doi.org/10.2139/ssrn.4628672

Fanta, N. – Horváth, R. (2024): *Artificial intelligence and central bank communication: the case of the ECB*. Applied Economics Letters, 32(18): 2600–2607. https://doi.org/10.1080/13504851.2024.2337318

Fulop, A. – Kocsis, Z. (2023): *News indices on country fundamentals*. Journal of Banking & Finance, 154, 106951. https://doi.org/10.1016/j.jbankfin.2023.106951

Gambacorta, L. – Kwon, B. – Park, T. – Patelli, P. – Zhu, S. (2024): *CB-LMs: language models for central banking.* BIS Working Paper No 1215. https://bis.org/publ/work1215.htm

Gábriel, P. – Pintér, K. (2006): *The effect of the MNB's communication on financial markets.* MNB Working Papers 2006/9, Magyar Nemzeti Bank. https://www.mnb.hu/letoltes/wp2006-9.pdf

Geiger, F. – Kanelis, D. – Lieberknecht, P. – Sola, D. (2025): *Monetary-Intelligent Language Agent (MILA)*. Deutsche Bundesbank Technical Paper No. 01/2025. https://hdl.handle.net/10419/316448

Gentzkow, M. – Kelly, B. – Taddy, M. (2019): *Text as Data*. Journal of Economic Literature, 57(3): 535–574. https://doi.org/10.1257/jel.20181020

Gorodnichenko, Y. – Pham, T. – Talavera, O. (2023): *The Voice of Monetary Policy*. American Economic Review, 113(2): 548–584. https://doi.org/10.1257/aer.20220129

Gössi, S. – Chen, Z. – Kim, W. – Bermeitinger, B. – Handschuh, S. (2023): *FinBERT-FOMC: Fine-Tuned FinBERT Model with Sentiment Focus Method for Enhancing Sentiment Analysis of FOMC Minutes.* ICAIF'23: Proceedings of the Fourth ACM International Conference on AI in Finance, pp. 357–364. https://doi.org/10.1145/3604237.3626843

Grandini, M. – Bagli, E. – Visani, G. (2020): *Metrics for Multi-Class Classification: An Overview*. arXiv preprint arXiv:2008.05756. https://doi.org/10.48550/arXiv.2008.05756

Gürkaynak, R.S. – Sack, B. – Swanson, E. (2005): *The Sensitivity of Long-Term Interest Rates to Economic News: Evidence and Implications for Macroeconomic Models*. American Economic Review, 95(1): 425–436. https://doi.org/10.1257/0002828053828446

Hansen, A.L. – Kazinnik, S. (2024): *Can ChatGPT Decipher Fedspeak?.* SSRN, April 11. https://doi.org/10.2139/ssrn.4399406

Hansen, S. – McMahon, M. (2016): *Shocking language: Understanding the macroeconomic effects of central bank communication*. Journal of International Economics, 99(S1): S114–S133. https://doi.org/10.1016/j.jinteco.2015.12.008

Hansen, S. – McMahon, M. – Prat, A. (2018): *Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach*. The Quarterly Journal of Economics, 133(2): 801–870. https://doi.org/10.1093/qje/qjx045

Horváth Dániel – Kálmán Péter – Kocsis Zalán – Ligeti Imre (2014): *Milyen tényezők mozgatják a hozamgörbét?.* MNB-Szemle, 2014. március: 28–39. https://www.mnb.hu/letoltes/horvath-kalman-kocsis-ligeti-1.pdf

Huang, A.H. – Wang, H. – Yang, Y. (2022): *FinBERT: A Large Language Model for Extracting Information from Financial Text.* Contemporary Accounting Research, 40(2): 806–841. https://doi.org/10.1111/1911-3846.12832

Kim, W. – Spörer, J. – Lee, C.L. – Handschuh, S. (2024): *Is Small Really Beautiful for Central Bank Communication? Evaluating Language Models for Finance: Llama-3-70B, GPT-4, FinBERT-FOMC, FinBERT, and VADER*. In: Proceedings of the 5th ACM International Conference on AI in Finance, pp. 626–633. https://doi.org/10.1145/3677052.3698675

Lewis, P. – Perez, E. – Piktus, A. – Petroni, F. – Karpukhin, V. – Goyal, N. et al. (2020): *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. Advances in Neural Information Processing Systems, 33: 9459–9474. https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

Liu, Y. – Ott, M. – Goyal, N. – Du, J. – Joshi, M. – Chen, D. et al. (2019): *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv preprint arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692

Liu, A. – Feng, B. – Xue, B. – Wang, B. – Wu, B. – Lu, C. et al. (2024): *DeepSeek-V3 Technical Report.* arXiv:2412.19437. https://doi.org/10.48550/arXiv.2412.19437

Mátrai-Pitz Mónika – Siket Bence (2023): *Reálgazdasági és inflációs tonalitásindexek.* Python-kód és dokumentáció. Mimeo.

Mikolov, T. – Chen, K. – Corrado, G. – Dean, J. (2013): *Efficient Estimation of Word Representations in Vector Space*. arXiv preprint arXiv:1301.3781. https://doi.org/10.48550/arXiv.1301.3781

Nagy Mohácsi, P. – Evdokimova, T. – Ponomarenko, O. – Ribakova, E. (2024): *Jegybanki politika és kommunikáció a feltörekvő országokban: A nagy felzárkózás.* Hitelintézeti Szemle, 23(1): 29–49. https://doi.org/10.25201/HSZ.23.1.29

Naszódi, A. – Csávás, C. – Erhart, S. – Felcser, D. (2016): *Which Aspects of Central Bank Transparency Matter? A Comprehensive Analysis of the Effect of Transparency of Survey Forecasts.* International Journal of Central Banking, 46(4): 147–192. https://www.ijcb.org/journal/ijcb16q4a4.pdf

Niţoi, M. – Pochea, M.-M. – Radu, Ş.-C. (2023): *Unveiling the sentiment behind central bank narratives: A novel deep learning index*. Journal of Behavioral and Experimental Finance, 38, 100809. https://doi.org/10.1016/j.jbef.2023.100809

Pennington, J. – Socher, R. – Manning, C.D. (2014): *Glove: Global Vectors for Word Representation*. In: Moschitti, A. – Pang, B. – Daelemans, W. (eds.): Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), October, pp. 1532–1543. https://doi.org/10.3115/v1/D14-1162

Peskoff, D. – Visokay, A. – Schulhoff, S. – Wachspress, B. – Blinder, A. – Stewart, B.M. (2024): *GPT Deciphering Fedspeak: Quantifying Dissent Among Hawks and Doves*. arXiv preprint arXiv:2407.19110. https://doi.org/10.48550/arXiv.2407.19110

Pfeifer, M. – Marohl, V.P (2023): *CentralBankRoBERTa: A Fine-Tuned Large Language Model for Central Bank Communications.* Journal of Finance and Data Science, 9, 100114. https://doi.org/10.1016/j.jfds.2023.100114

Picault, M. – Renault, T. (2017): *Words are not all created equal: A new measure of ECB communication*. Journal of International Money and Finance, 79: 136–156. https://doi.org/10.1016/j.jimonfin.2017.09.005

Poole, W. – Rasche, R.H. – Thornton, D.L. (2002): *Market Anticipations of Monetary Policy Actions*. Review-Federal Reserve Bank of Saint Louis, 84(4): 65–94. http://doi.org/10.20955/r.84.65-94

Romer, C.D. – Romer, D.H. (2000): *Federal Reserve Information and the Behavior of Interest Rates*. American Economic Review, 90(3): 429–457. http://doi.org/10.1257/aer.90.3.429

Seelajaroen, R. – Budsaratragoon, P. – Jitmaneeroj, B. (2020): *Do monetary policy transparency and central bank communication reduce interest rate disagreement?*. Journal of Forecasting, 39(3): 368–393. http://doi.org/10.1002/for.2631

Takahashi, K. – Yamamoto, K. – Kuchiba, A. – Koyama, T. (2022): *Confidence interval for micro-averaged F_1 and macro-averaged F_1 scores*. Applied Intelligence, 52(5): 4961–4972. https://doi.org/10.1007/s10489-021-02635-5

Thorsrud, L.A. (2018): *Words are the New Numbers: A Newsy Coincident Index of the Business Cycle.* Journal of Business & Economic Statistics, 38(2): 393–409. https://doi.org/10.1080/07350015.2018.1506344

Tobback, E. – Nardelli, S. – Martens, D. (2017): *Between hawks and doves: measuring central bank communication.* ECB Working Paper No. 2085. https://doi.org/10.2139/ssrn.2997481

Touvron, H. – Lavril, T. – Izacard, G. – Martinet, X. – Lachaux, M-A. – Lacroix, T. et al. (2023): *Llama: Open and Efficient Foundation Language Models*. arXiv preprint arXiv:2302.13971. https://doi.org/10.48550/arXiv.2302.13971

Vaswani, A. – Shazeer, N. – Parmar, N. – Uszkoreit, J. – Jones, L. – Gomez, A.N. et al. (2017): *Attention Is All You Need*. Advances in Neural Information Processing Systems, 30. https://doi.org/10.48550/arXiv.1706.03762

Wei, J. – Wang, X. – Schuurmans, D. – Bosma, M. – Ichter, B. – Xia, F. et al. (2022): *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. Advances in Neural Information Processing Systems, 35, 24824–24837. https://doi.org/10.48550/arXiv.2201.11903

Yao, R. – Chai, Q. – Yao, J. – Li, S. – Chen, J. – Zhang, Q. et al. (2025): *Interpreting Fedspeak with Confidence: A LLM-Based Uncertainty-Aware Framework Guided by Monetary Policy Transmission Paths*. arXiv preprint arXiv:2508.08001. https://doi.org/10.48550/arXiv.2508.08001

Yao, S. – Zhao, J. – Yu, D. – Du, N. – Shafran, I. – Narasimhan, K. et al. (2023): *ReAct: Synergizing Reasoning and Acting in Language Models*. In: International Conference on Learning Representations (ICLR), January. https://doi.org/10.48550/arXiv.2210.03629

Függelék

5. táblázat
Részletes eredmények a tesztmintán (REAL-fundamentum, kategóriánkénti megfigyelések)

	Pozitív			Semleges			Negatív			Nem fundamentum		
	TP	FP	FN	TP	FP	FN	TP	FP	FN	TP	FP	FN
Evdokimova et al. (2023)	160	115	62	0	2	75	122	74	91	1 182	162	125
Mátrai-Pitz – Siket (2023)	127	78	95	7	36	68	96	44	117	1 227	202	80
Fulop – Kocsis (2023)	112	82	110	8	16	67	87	32	126	1 242	238	65
Picault – Renault (2017)	114	370	108	0	11	75	143	618	70	488	73	819
Huang et al. (2022)	199	26	23	41	17	34	197	35	16	1 270	32	37
Gössi et al. (2023)	199	33	23	39	19	36	195	58	18	1 248	26	59
Pfeifer – Marohl (2023)	197	30	25	39	21	36	200	51	13	1 257	22	50
Gambacorta et al. (2024)	206	43	16	39	16	36	195	45	18	1 249	24	58
GPT-4.1-nano	184	163	38	17	101	58	127	48	86	1 147	30	160
GPT-5-nano	133	45	89	9	8	66	157	26	56	1 276	163	31
GPT-4.1-mini	187	66	35	22	21	53	167	40	46	1 251	63	56
GPT-5-mini	207	62	15	23	34	52	192	48	21	1 223	28	84
GPT-4.1	183	56	39	19	21	56	168	52	45	1 242	76	65
GPT-5	206	50	16	26	15	49	184	42	29	1 250	44	57

Megjegyzés: A táblázat a 3.1. és 3.2. alfejezetek eredményeinek részletesebb, esetszám-alapú közlése. A REAL-fundamentum esetében értékkategóriánként (pozitív, semleges, negatív, nem fundamentum) mutatja be az pozitív találatok (TP: true positive), az elsőfajú hiba (FP: false positive, a „téves riasztás" esete), és a másodfajú hiba (FN: false negative, az „elmulasztott találat" esete) elemszámait. A minta 1 817 elemet tartalmaz (ez a 2 384 elemű tesztmintából már kiszűri azokat az eseteket, amikor a humán annotációk nem egyeznek). A könnyebb értelmezés érdekében TP esetében a magasabb elemszámok sötétebb zöld árnyalatot, az FP és FN esetén magasabb elemszámok sötétebb piros árnyalatot kapnak.

Forrás: A szerzők számításai

6. táblázat

Részletes eredmények a tesztmintán (INFLA-fundamentum, kategóriánkénti megfigyelések)

	Pozitív			Semleges			Negatív			Nem fundamentum		
	TP	FP	FN	TP	FP	FN	TP	FP	FN	TP	FP	FN
Evdokimova et al. (2023)	118	97	58	0	0	107	89	55	54	1 361	151	84
Mátrai-Pitz – Siket (2023)	104	52	72	5	27	102	93	46	50	1 395	149	50
Fulop – Kocsis (2023)	99	51	77	4	15	103	77	36	66	1 407	182	38
Picault – Renault (2017)	–	–	–	–	–	–	–	–	–	–	–	–
Huang et al. (2022)	157	48	19	77	40	30	119	32	24	1 389	9	56
Gössi et al. (2023)	154	53	22	62	34	45	123	43	20	1 391	11	54
Pfeifer – Marohl (2023)	157	46	19	73	34	34	120	27	23	1 405	9	40
Gambacorta et al. (2024)	155	52	21	66	33	41	121	32	22	1 400	12	45
GPT-4.1-nano	127	73	49	55	121	52	71	32	72	1 359	33	86
GPT-5-nano	105	21	71	8	19	99	87	21	56	1 425	185	20
GPT-4.1-mini	146	89	30	31	31	76	110	46	33	1 371	47	74
GPT-5-mini	150	62	26	38	39	69	126	58	17	1 368	30	77
GPT-4.1	124	49	52	31	31	76	102	49	41	1 394	91	51
GPT-5	147	45	29	40	22	67	129	61	14	1 393	34	52

Megjegyzés: A táblázat a 3.1. és 3.2. alfejezetek eredményeinek részletesebb, esetszám-alapú közlése. Az INFLA-fundamentum esetében értékkategóriánként (pozitív, semleges, negatív, nem fundamentum) mutatja be a pozitív találatok (TP: true positive), az elsőfajú hiba (FP: false positive, a „téves riasztás" esete), és a másodfajú hiba (FN: false negative, az „elmulasztott találat" esete) elemszámait. A minta 1 871 elemet tartalmaz (ez a 2 384 elemű tesztmintából már kiszűri azokat az eseteket, amikor a humán annotációk nem egyeznek). A könnyebb értelmezés érdekében TP esetében a magasabb elemszámok sötétebb zöld árnyalatot, az FP és FN esetén magasabb elemszámok sötétebb piros árnyalatot kapnak.

Forrás: A szerzők számításai

7. táblázat

Részletes eredmények a tesztmintán (MPOL-fundamentum, kategóriánkénti megfigyelések)

	Pozitív			Semleges			Negatív			Nem fundamentum		
	TP	FP	FN	TP	FP	FN	TP	FP	FN	TP	FP	FN
Evdokimova et al. (2023)	48	164	64	1	4	88	61	190	90	1 202	165	281
Mátrai-Pitz – Siket (2023)	-	-	-	-	-	-	-	-	-	-	-	-
Fulop – Kocsis (2023)	7	50	105	1	2	88	0	38	151	1 480	257	3
Picault – Renault (2017)	13	219	99	1	10	88	112	832	39	571	77	912
Huang et al. (2022)	92	29	20	47	33	42	132	39	19	1 448	15	35
Gössi et al. (2023)	79	20	33	49	35	40	123	55	28	1 446	28	37
Pfeifer – Marohl (2023)	87	31	25	46	33	43	132	44	19	1 448	14	35
Gambacorta et al. (2024)	75	16	37	40	21	49	129	61	22	1 457	36	26
GPT-4.1-nano	40	27	72	48	185	41	65	111	86	1 319	40	164
GPT-5-nano	58	34	54	9	13	80	85	15	66	1 441	180	42
GPT-4.1-mini	95	47	17	31	63	58	129	36	22	1 391	43	92
GPT-5-mini	99	45	13	21	24	68	131	46	20	1 422	47	61
GPT-4.1	92	47	20	22	25	67	132	54	19	1 422	41	61
GPT-5	96	43	16	25	20	64	132	43	19	1 424	52	59

Megjegyzés: A táblázat a 3.1. és 3.2. alfejezetek eredményeinek részletesebb, esetszám-alapú közlése. Az MPOL-fundamentum esetében értékkategóriánként (pozitív, semleges, negatív, nem fundamentum) mutatja be a pozitív találatok (TP: true positive), az elsőfajú hiba (FP: false positive, a „téves riasztás" esete), és a másodfajú hiba (FN: false negative, az „elmulasztott találat" esete) elemszámait. A minta 1 835 elemet tartalmaz (ez a 2 384 elemű tesztmintából már kiszűri azokat az eseteket amikor a humán annotációk nem egyeznek). A könnyebb értelmezés érdekében TP esetében a magasabb elemszámok sötétebb zöld árnyalatot, az FP és FN esetén magasabb elemszámok sötétebb piros árnyalatot kapnak.

Forrás: A szerzők számításai