TDK-dolgozat

Balogh Etele

LLM-alapú, struktúrálatlan orvosi leletekből történő adatkinyerési módszerek összehasonlítása

Balogh Etele II. évf. Programtervező Informatikus MSc
Témavezetők: Dr. Kicsi András, Dr. Vidács László
Szegedi Tudományegyetem, Természettudományi és Informatikai Kar,
Szoftverfejlesztés tanszék

Tartalomjegyzék

1. Absztrakt **4**

2. Bevezetés **5**

2.1. Saját kontribúció 6

3. Háttér **7**

3.1. Adat 10

3.1.1. Tanító adat 10

3.1.2. Kiértékelési adat 11

3.2. Kimeneti formátumok 11

3.3. Módszerek 12

3.3.1. Finomhangolt BERT 12

3.3.2. GPT 14

3.4. Kiértékelési szempontok 16

3.4.1. Mérőszámok 16

3.4.2. Szintek 17

4. Kísérletek **19**

4.1. Eredmények 19

4.2. Eredmények értelmezése 21

4.2.1. Módszerek felhasználhatósága 23

5. Összefoglalás **25**

5.1. Továbblépési lehetőségek 25

6. Köszönetnyílvánítás **26**

7. Melléklet **27**

Hivatkozások **28**

1. Absztrakt

Az elmúlt években a számítástudományok területét árvízként lepték el a nagy nyelvi modellek. Képességeiket sok területen próbálják hasznosítani, általánosságban egyre inkább növekvő sikerrel. Ilyen terület például a nyelvészet vagy az oktatás. Kutatásom során az orvostudomány, azon belül a radiológiai leletezés területére koncentráltam. A Röntgen és MR felvételek kiértékelése során nagy mennyiségű szöveges és képi adat keletkezik, ami a radiológiai vizsgálatok fő kimeneti terméke, és amit napjainkban emberi erőbefektetéssel dolgoznak fel, ami egy költséges erőforrás.

A dolgozatomban bemutatott munkánkat egy korábbi, munkatársaim által írt cikk eredményeire alapoztuk. Célunk az volt, hogy összehasonlítást készítsünk a cikkben használt, publikusan elérhető, Generative Pre-trained Transformerre (GPT-re) alapuló módszer, és egy általunk készített, Bidirectional Encoder Representations from Transformers-et (BERT-et), egyszerű ontológiákat, és szabály-alapú megoldásokat használó módszer eredményességét tekintve a leletek automatizált értelmezésében. A különböző módszertanok egy radiológus által annotált adathalmazon kerültek kiértékelésre 56 valódi, angol nyelvű leleten. Az eredményeket az értelmezés több szintjén, több szempont alapján hasonlítottuk össze. A dolgozatban bemutatott kutatás fő kérdésköre, hogy egy általános nagy nyelvi modell mennyivel marad el, vagy teljesíti túl a célzottan radiológiai leletekre kihegyezett módszert. Következtetéseink során figyelembe vettük a két megközelítés előnyeit és hátrányait is.

Eredményeink azt mutatják, hogy a két módszer megközelítőleg hasonló pontossággal teljesít. Mindkét módszer hatékony a leletek kiértékelésében, ami potenciálisan elősegítheti a radiológiai szöveges leletek gépi feldolgozásának területét. Megállapítottuk, hogy a két módszer a különböző részfeladatokon eltérően teljesít, méréseink szerint a GPT jobb a kapcsolatok felismerésében, míg az erre előtanított és finomhangolt BERT pontosabban végzi a tokenek osztályozását.

2. Bevezetés

Az elmúlt években a számítástudományok területén óriási érdeklődésre tettek szert a nagy nyelvi modellek. Laikusok körében népszerűségét köszönhetik annak, hogy különböző online eszközök által könnyen elérhetőek (például ChatGPT [1], Gemini[2]), kényelmesen és könnyen használhatóak, a belőlük kinyerhető tudásanyag pedig vetekedik a legnagyobb internetes tudásbázisokban fellelhetőkkel. Képességeiket több-kevesebb sikerrel hasznosítják az oktatás és nyelvészet területén is.

A nagy nyelvi modellek, Large Language Modells (LLM), nagymennyiségű szöveges adaton tanított mélytanuló mesterséges neuron hálók. Képesek szöveg kiegészítést, egy téma alapján szöveg generálást végezni, vagy egy adott szövegre vonatkozó kérdéseket megválaszolni.

A dolgozatom készítése során az emberi gerincről készült Röntgen/MR felvételek feldolgozása során keletkező nagy mennyiségű szöveges orvosi leletek feldolgozására koncentráltam. A sajátmodelljeink is erre a domainre lettek finomhangolva.

A XXI. századra a kutatók és ipari szakemberek már elkezdtek különböző informatikai eszközöket is bevezetni, hogy a leletfeldolgozási folyamatot felgyorsítsák, mint például képi szegmentáció [3], vagy a strukturálatlan szöveges leletekből való adatkinyerésre alkalmas algoritmusok használata. A dolgozatban taglalt kutatói munkám alapját az a feltevés képezte, hogy a fentebb említett LLM-ek használata előre tudná-e lendíteni ezen tudományágat, vagy a jelenlegi munkaterhet tudná-e csökkenteni.

Munkámat egy korábbi, munkatársaim által, a 2024-es év során közölt cikk és az abban taglalt eredmények [4] felhasználásával végeztem. Célom az volt, hogy összehasonlítást készítsek a cikkben használt Generative Pre-trained Transformerre (GPT-re) [5] alapuló módszertan és egy Bidirectional Encoder Representations from Transformers-t (BERT-et) [6] használó módszerek eredményességét tekintve. A munka során alapos elemzés alá vetettem a két módszertan entitás-címkézési, és kapcsolatfelismerési képességeit.

A különböző módszertanok egy radiológus által annotált adathalmazon kerültek kiértékelésre. A halmaz 56 különböző, valódi, angol nyelvű radiológiai

leletet tartalmazott, amik véletlenszerűen lettek kiválasztva a MIMIC III [7] és MTSamples[8] adatbázisokból. A kiértékelést manuálisan végeztem a radiológus által annotált leletekkel összehasonlítva a kimeneteket. A kézi összehasonlításra azért volt szükség, mert egyfelől a két módszer más formátumú kimenetet ad még szkriptekkel való formázás után is, illetve mert a kiértékelést több szinten kellett elvégezni amiben az egyes módszerek által vétett kezdeti hibákat nem akartuk továbbvinni a későbbi szintekre is.

A konklúzió levonása során figyelembe vettem olyan tényezőket, mint a különböző LLM-ek végfelhasználói használhatóságának lehetőségei, és egyéb előnyeik és hátrányaik, amelyek specifikusan rájuk vonatkoznak.

2.1. Saját kontribúció

Munkánk során az én egyéni hozzájárulásom többek között az volt, hogy a BERT-re épülő csővezeték főbb alfeladataihoz, mint a token-osztályozás és reláció-kinyerés, új modelleket terveztem és tanítottam a BioBERT [9] alap modellre építve. A két modellhez a tanító, validációs és teszt adathalmaz szintén a MIMIC III és MTSamples-ből lett leválogatva, ahol átfedés csak a teszthalmazok között volt.

A BERT-alapú csővezetéket egy korábbi projekt keretein belül már felhasználtuk, ám a pontosabb összehasonlíthatósághoz ennek újratervezésére és újabb finomhangolására volt szükség a GPT-vel megegyező entitások használatához. Ezt az új finomhangolást elvégeztem, és kiegészítettem egy egyedi utófeldolgozás fázissal és testre szabtam a különböző fázisait, hogy a kívánt formában adja meg a kimenetet.

A GPT-kiértékelő csővezeték elkészítése során az Azure OpenAI API-t[1] használtuk, amihez a már említett, munkatársaim által publikált cikkben megjelent promptot használtam és az ahhoz készült Python szkriptet használtam.

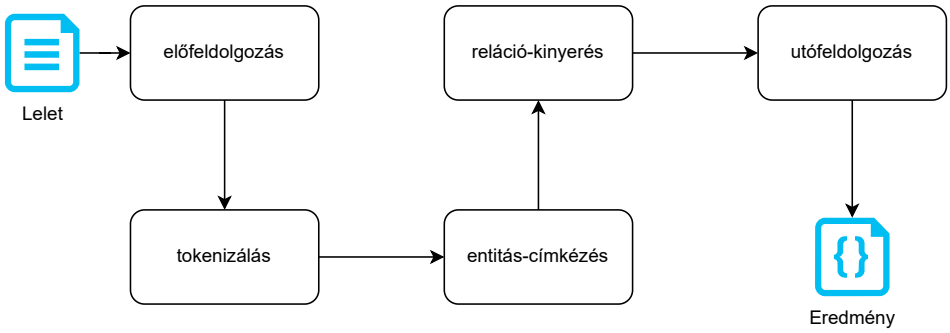
Az eredmények kiértékelését egy előre meghatározott szempontrendszer mentén, manuálisan végeztem az általam kiválasztott teszthalmazon. Az összehasonlításhoz szükséges kísérleteket én terveztem meg, ami felhasználta az általam készített eljárás alapján átalakított kimeneteket.

[1]A MIMIC-III adatbázis felhasználási feltételeiken megfelelően.

Végeztem kísérleteket arra is, hogy a saját módszerünk kimenetét átalakítsam a GPT álta előállított kimenetre.

A kutatói munka során használt módszerek mindegyike a múltban kutatva és publikálva volt. Az eredményeik összehasonlításához szükséges kísérletek tervezése és implementációja az én munkám volt.

3. Háttér



1. ábra. **BERT** csővezeték

Mind a GPT, mind a BERT a nagy nyelvi modellek családjába tartozik. Mindkét model a transformer architektúra valamilyen variánsára épült. A GPT-t nagy mennyiségű címkézetlen adathalmazon tanították. A BERT-nél önfelügyelt, token-maszkolásos tanítást alkalmaztak. A modellek eredeti tanító adata a BookCorpus [10], WebText [11] valamint English Wikipedia [12].

Mindkét modellt lehetséges domain-specifikus feladatokra finomhangolni, a BERT esetében a csak encodert alkalmazó architekrúra lehetővé teszi, hogy az utolsó réteg lecserélésével meg lehessen ezt tenni. Ilyen specifikus modellekre példa a BioGPT [13], ami az orvosi szövegeken tanított modell, a ChatGPT, ami egy általános célú chat-bot, a fentebb említett BioBERT vagy a SciBERT orvosi és tudományos domainre [14].

A nagy nyelvi modellekkel történő struktúrált adatkinyerés, orvosi szövegfeldolgozás aktívan kutatott terület. Xieling Chen és társai publikációja [15] alapján **2007** és **2016** között *1405* természetes nyelvfeldolgozással (NLP-vel)

támogatott orvosi domainel foglalkozó publikáció jelent meg. Az eredmények alapján a publikációk éves átlagos növekedése *18,39%*-volt. Kutatásuk szerint a fő kutatott területek közé tartoztak a terminológiabányászat, információkinyerés, szöveg-osztályozás stb.

Matthew C. Chen és társai 2017-es cikkükben mély tanuló konvolúciós neuron hálóval (CNN) végeztek kísérleteket a szabadszöveges radiológiai leletek oszályozására [16]. Mellkasi computed tomography (CT) felvételekkel dolgoztak, amikben tüdőembóliát kerestek és 3 osztályba sorolták a leleteket: jelenlét, krónikusság és helyzet. Kutatásuk során, a CNN használatával sikerült *99,00%*-os pontosságot és *93,80%*-as F1-mértéket elérniük.

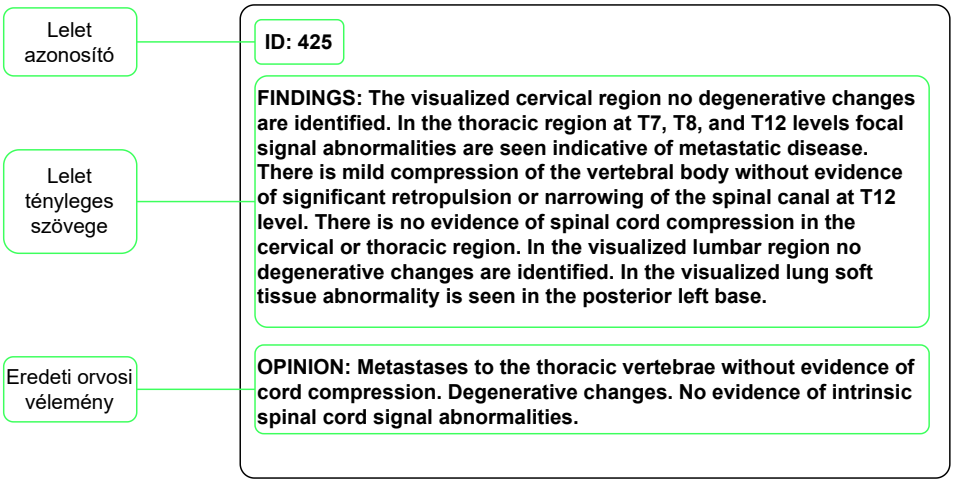
A dolgozatban taglalt összehasonlításon kívül más kutatások is célozták már a GPT és BERT teljesítményének összehasonlítását. Gutierrez és társai [17] 2022-es publikációjukban még a GPT-3-hoz mérve a BERT-alapú RoBERTa-large-hoz [18] képest (de két másik BERT variánshoz képest is) 100 tanítópéldán finomhangolva angol nyelvű orvosi entitás- és reláció-kinyerésében arra jutottak, hogy a GPT jelentősen elmarad a BERT eredményeitől. Itt betegségeket és vegyületeket, valamint azok egymáshoz való kapcsolatait mérték. A jobban teljesítő RoBERTa-large az entitás-kinyerésben átlagosan *67,2%*-os, míg a relációk kinyerésében *47,4%*-os F-mértékkel teljesített esetükben.

BERT-re és GPT-re alapuló rendszerek alkalmazásával már nagyban folynak kutatások világszerte, amelyek ígéretes eredményeket képesek felmutatni, a modellek precizitása valamint az alkalmazhatóságuk szempontjából. Bo Guo és társai BERT-re és feltételes valószínűségi mezőkre (CRF) épülő megoldása leletszövegekben keresett betegségeket [19], tüneteket, diagnózisokat, gyógyszereket, és vizsgálati eredményeket, valamint gyógyszer-dózis, gyógyszer-betegség, és betegség-tünet kapcsolatokat. Az entitás-kinyerési feladaton *91,34%-94,63%* közötti F-mértéket, míg a kapcsolat-kinyerési feladaton *87,53%-94,92%* közötti F-mértéket értek el különböző vizsgált adathalmazaikon. Gams és társai [20] GPT-4-es képességeket integráltak egy orvosi információs platformba, amelyen keresztül a páciensek egészségügyi témakörben, chat-bottal kommunikálva informálódhatnak, az orvosi tudást a platform ellenőrzött egészségügyi információból adja a rendszer, személyes adatokhoz

vagy leletekhez nincs hozzáférése. Vizsgálataik szerint a platform információi segítenek a hallucináció minimalizálásában, illetve kvalitatív elemzésük során a módszert orvosokkal kipróbálva magas pontosságot állapítottak meg.

Chow és társai [21] is végeztek kutatásokat ChatGPT alapú orvosi chat-bot témában, azt találták, hogy az elképzelés ígéretes, viszont az éles környezetben történő használatát befolyásolja a pontossága és megbízhatósága, valamint az, hogy nem képes valódi orvosi szakértelem reprodukálására. Meglátásaik hasonlóak a Chakraborty és társai [22] által végzet kutatás eredényeihez, ahol megállapították, hogy a ChatGPT és a chat-botok általánosságban, csak részben tudják helyettesíteni az orvosi munkát, tanácsadásra alkalmasak lehetnek, de megfelelően diagnosztizálni nem tudnak. Thirunavukarasu és társai folytattak kutatásokat olyan irányban, hogy egy LLM modellre alapuló chat-botot tudjanak kínálni a felhasználók számára [23], maga az irány járható, az eredmények felemásak viszont kimondottan ígéretesek, ellenben továbbra is fennáll a fentebb kiemelt jogi kérdés, ami az ilyen ipari projektek kivitelezését erősen megnehezíti.

A modern BERT modellek orvosi domainen történő adatkinyerésére is voltak kutatások, mint a Sushil és társai által publikált [24] eredmények, ahol megállapították, hogy a felügyeletnélküli adatkinyerés még túl komplex ahhoz, hogy ezek a megoldások teljes mértékben meg tudják oldani.



2. ábra. Példa egy angol nyelvű lelet és véleményre

	Tanító	*Validációs*	*Teszt*
Other	**3139**	**571**	**940**
TestreszElvaltozas	**4352**	**529**	**1183**
Tagadva	**643**	**71**	**167**

1. táblázat. Az R-BERT finomhangolásához használt adatbázis felépítése, mondatokra bontva

A kutatásunk során használt lelethalmaz valós, angol nyelvű, anonimizált leletekből állt össze. Ezek a leletek 3 fő részből álltak fel (lásd a 2. ábrán), mindegyik lelet rendelkezik egy egyedi azonosítóval, ami a "forrás adatbázis" + "szám" (lehetett nullával feltöltött, egyjegyű szám is). Az azonosítón felül mindegyik lelet tartalmaz egy "findings" mezőt, ami a lelet tényleges szövegét tartalmazta, valamint egy "opinion" részt, ami a leletet eredetileg lejegyző orvosnak a "findings" mezőre alapozottan felállított véleménye.

3.1. Adat

Kutatómunkák során a már fent említett MIMIC-III és MTSamples adatbázisokat használtuk, mindkét adatbázis valós, angol nyelvű, anonimizált orvosi leleteket tartalmaz. Munkánk során *150* MIMIC leletet és *53* MTSamples leletet használtunk, ennek a halmaznak egy részét (*56* leletet) választottuk ki a manuális kiértékeléshez.

3.1.1. Tanító adat

Az entitás-címkézésre finomhangolt BERT alapú modell adata *2806* mondat címkézett tokenjeiből állt, amiket egy beépített függvénnyel választott szét a rendszer tanító, validációs és teszt halmazokra, így kaptunk *1963* mondatnyi tanító, *281* mondatnyi validációs és *562* mondatnyi teszt halmazt. A reláció-kinyerő R-BERT[25] modell adataira már általunk lett előkészítve az adathalmaz, aminek a felépítése az 1. táblázatról leolvasható. A táblázatban feltüntetett számok az egyes kapcsolat típusokat jelképezik. A címkék úgy vannak megadva, hogy például: "Tagadva(e1,e2)", "Tagadva(e2,e1)" a kapcsolat iránya a táblázatban nincsen elkülönítve, ám egy kapcsolat egyszer

	Felhasznált	*Entitások*	*Kapcsolatok*
MIMC-III	**50**	**1270**	**910**
MTSamples	**6**	**221**	**160**

2. táblázat. Kiértékeléshez használt halmaz összetétele

fordult elő az adatokban mindig.

3.1.2. Kiértékelési adat

A kiértékelési halmaz olyan leleteket tartalmazott, amik a BERT alapú modellek tanítása során, legfeljebb csak a teszt halmazban szerepeltek. Ez a halmaz a 2. táblázatban látható módon *50* MIMIC-III és *6* MTSamples leletet tartalmazott, amikhez rendelkezésre álltak az orvosi annotációk.

3.2. Kimeneti formátumok

XXXX-00001

Finding:The cervical cord appears normal in its size and signal characteristics. The C2-3 and C3-4 discs are degenerated. At C2-3 there is disc desiccation with a posterior central disc herniation. There is some mild bulging of the C3-4 disc. Neither level demonstrates central or neural foraminal narrowing. At C4-5 and C5-6 there is no recurrent central or neural foraminal narrowing. At C6-7 there is mild bilateral bony neural foraminal narrowing without central canal compromise. The C7-T1 level appears unremarkable. MRI of the brain is broadly within normal limits.

EXTRASPINAL SENTENCES:
- MRI of the brain is broadly within normal limits.

NON DEGENERATIVE SENTENCES:
- The cervical cord appears normal in its size and signal characteristics.
- The C7-T1 level appears unremarkable.

DEGENERATIVE SENTENCES:
- The C2-3 and C3-4 discs are degenerated.
- level: C2-3, C3-4
- anatomy: disc
- degeneration: degenerated
- At C2-3 there is disc desiccation with a posterior central disc herniation.
- level: C2-3
- anatomy: disc
- degeneration: desiccation
****
- level: C2-3
- anatomy: disc
- degeneration: herniation
...

3. ábra. A GPT-alapú módszer kimeneti formátuma

Kutatómunkánk során több fajta kimeneti formátumot is kialakítottunk.

A GPT kimenetként hivatkozott formátum felépítése a 3. ábrán látható módon került összeállításra, ez a formátum tartalmazza az eredeti bemeneti szöveget, aztán mondatonkénti csoportosításban az alapján, hogy az adott mondat gerincen kívüli témáról szól, a gerincről szól, viszont nem degeneratív elváltozást ír le (pl. normális állapot, ép porckorong, vagy megtartott ív), vagy degeneratív elváltozást ír le a gerincen.

Az orvosi annotáció és a BERT kimenete soronként tartalmazza először a feldolgozás során kinyert entitásokat, majd az entitások között azonosított kapcsolatokat. Az entitásokat jelölő sorok lényegi részei az ábrán látható módon a következő felépítést követik:
"*Azonosító* \t *Címke* \t **start index _ vég index** \t *szöveg részlet*",
ahol a "_" a szóközöket, míg a "\t" tabulátorokat jelöl. A kapcsolatok jelölése tartalmaz szintén egy azonosítót, egy címkét, a kapcsolatban résztvevő két entitás azonosítóját. Ezt a formátumot a brat annotációs eszköz [26] kiemenetére építettük, amely eleve ilyen formában tárolja az adatokat, .ann kiterjesztésű fájlokban.

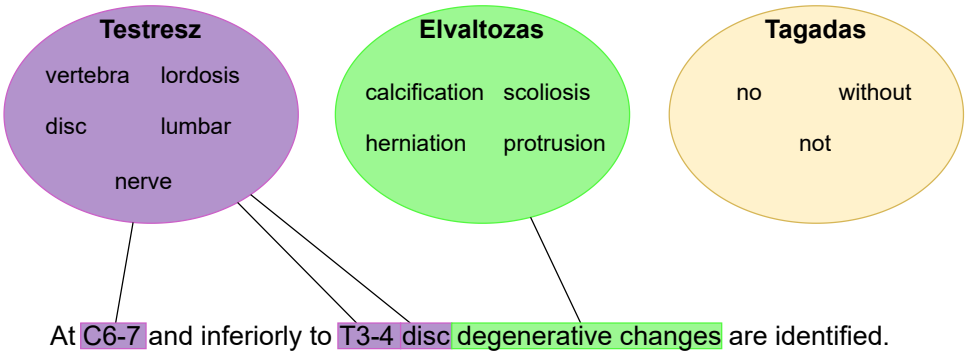
3.3. Módszerek

Kutatásunk során két módszert vizsgáltunk orvosi szöveges leletfeldolgozás szempontjából. Mindkét módszer nagy nyelvi modellek köré épül. A kutatás alapjául szolgáló cikkben megalapozott GPT-s kiértékelő, valamint egy, a korábbi kutatási projekt keretein belül elkészített [27] és általam finomhangolt, specifikus domainekre tanított, BERT modelt felhasználó megoldás.

3.3.1. Finomhangolt BERT

A BERT használata elterjedt a számítógépes nyelvészet területén, az alap előtanított modellt egyszerű akár csak egy réteg hozzáadásával finomhangolni a kívánt feladat elvégzésére.

A saját készítésű, BERT-re alapuló lelet-értelmező megoldást csővezetékesítettük (lásd az 1. ábrát), amelynek lényege az, hogy a különböző lépések kimenetei közvetlenül be vannak kötve a következő lépés bemenetére.



4. ábra. Példa az entitás-címkézési feladatra, a különböző entitásokra néhány példa megadásával, és egy egyszerű mondaton belüli hozzárendelésével.

A csővezeték bemenetként, a BERT a GPT-től eltérően az egész leletet megkapta, amiből utána kinyerte a kívánt *findings* mezőt, ezen a szövegen az ábrán látható előfeldolgozás fázis során elvégzett néhány további lépést, ezek a lépések tartalmazzák a beérkező lelet felbontását azonosító, lelet szöveg és vélemény mezőkre. A lelet szövegén az előfeldolgozás elvégez egy szűrést, hogy a lehetséges karater kódolási hibákat javítsa, majd ezt a javított szöveget felbontja mondatokra, hogy a kiértékelés után össze tudjon állítani egy, a GPT kimenetével megegyező kimenetet az utófeldolgozási fázisban.

Az előfeldolgozott leletszöveget adtuk át először egy *Spacy* tokenizálónak [28], amit utána a token osztályozó követett.

A token-osztályozási feladatot egy, az erre a feladatra finomhangolt BERT modell végezte, ami 3 különböző címkét figyelt: *Testresz*, *Evaltozas* és *Other*, illetve ezeknek Beginning és Inside változatait a több tokenből álló kifejezések kezelésére. Ilyen típusú feladatokra, valamit a reláció kinyerésre már régóta vizsgálják a BERT-alapú megoldásokat [19].

A kapcsolat-felismerést szintén egy finomhangolt BERT-en alapuló, R-BERT modell végezte. Ez felhasználta a token-osztályozás lépésben, valamint a szabály-alapú módszer által meghatározott *Tagadas* címkéket. A modell 3 különböző kapcsolat felismerésére képes: "TestreszElvaltozasa", "Tagadva", "Other" ahol az utóbbi azt jelöli, hogy nincsen kapcsolat. Szabály-alapon továbbá képes "TestreszReszei" relációt detektálni a rendszer, ami a "TestreszElvaltozasa" kapcsolatok és a szöveg-környezet alapján hoz döntést.

Az R-BERT egy BERT-en alapuló modell, amit kifejezetten a kapcsolatok felismerésére készítettek. Itt a modell bemenetként egy mondatot kap, melyben különböző karakterekkel ($, #) van megjelölve a két kapcsolatban álló szó. A kijelölt entitások több tokenre vannak felbontva a BERT kimenetében, így az egy entitáshoz tartozó kimeneteket először átlagolják, majd tanh aktivációt követően mindkét kifejezés egy teljesen kapcsolt rétegen megy keresztül. Az első token ([CLS]) végső rejtett kimenetére is ugyanígy alkalmaznak egy aktivációs függvényt, majd egy teljesen kapcsolt réteget. Az így kapott három kimenetet összefűzik, mely egy teljesen kapcsolt és egy softmax rétegen megy keresztül. Az R-BERT publikálói szerint jelentősen jobban (89.25%-os F-mérték) teljesített kapcsolatok felismerésében a korábban a feladatra népszerű Bi-LSTM és CNN módszereknél (attention-t is felhasználó megoldásokban ezek 85,2% és 88,0% F-mértéket értek el, ebben a sorrendben)

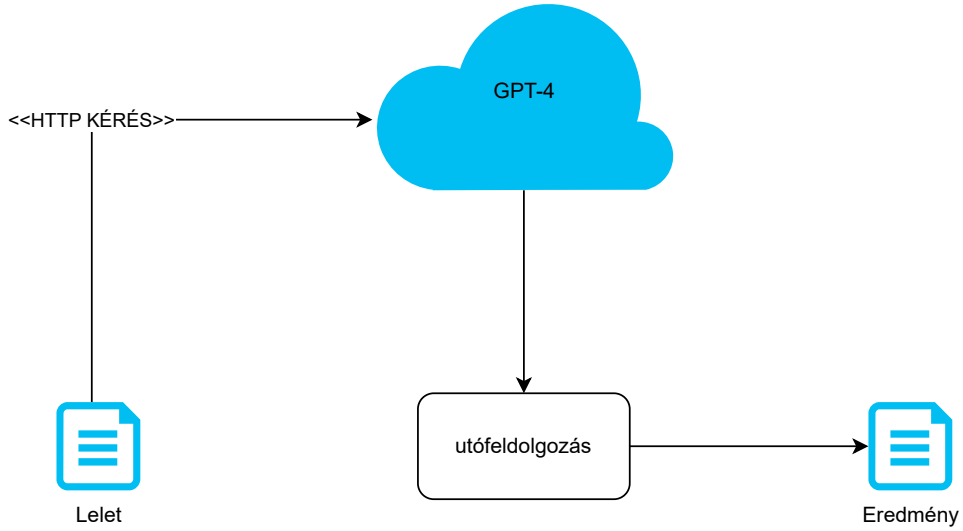
A csővezeték utolsó lépéseként az elkészült annotációs fájlt átalakítottuk egy olyan formátumra, ami az eredeti, GPT-vel struktturálástt végző cikkben is volt, ami lehetővé tette a két módszer közvetlen összehasonlítását. A végső kimenet tartalmazta mind az orvosi annotációs fájl mintájára összeállított eredményt, mind pedig a GPT kimenetének formátumára átalakított változatot.

3.3.2. GPT

A fejletebb GPT alapú kutatások az orvosi szövegfeldolgozás területén még újkeletűek, maga a lehetőség nemrég nyílt meg a nagy nyilvánosság számára. Ennek ellenére már számos kutatás és publikáció készült arról, hogy milyen módokon lehet olyan promptokat, utasítássorozatokat összeállítani, hogy prompt, vagy kontextus alapú tanulást el lehessen érni [29] a kívánt feladatok teljesítésére. A publikált eredmények alapján a GPT modell képes volt kimagasló eredményeket elérni few-shot vagy akár zero-shot kontextusbeli tanulással is.

A GPT futtatások alapját a munkatársaim korábbi cikkében publikált prompt[2] adta, ami utasításokat tartalmaz arra, hogy a beérkező lelet szövegét

[2]A prompt megtekinthető a csatolmányban található GitHub repozitóriumban.



5. ábra. **GPT** csővezeték

milyen lépésekben, hogyan kell feldolgozni. A lényegét tekintve ez is egy csővezeték, viszont a nagy része absztrahálva van a GPT-4 rendszerében, ami feketedoboz-szerűen elvégzi a különböző, egymást követő lépéseket.

A prompt 129 sorból áll. 6 fő lépést ír le, minden megfogalmazott feldolgozási lépéshez tartalmaz egy mintát, hogy az adott fázis végén hogyan kellene reprezentálnia az adatot. Tartalmaz a végén továbbá egy teljes példabemenetet és a hozzá elvárt példakimenetet, valamint egy utasítást, hogy a kimenet, csak az utolsó lépés eredményét tartalmazza.

Mindegyik fázis mondatszinten vizsgálta a szöveget. Első lépésben a mondatokat egyesével osztályozta, hogy a tartalmuk a gerincoszlopot érintette-e, vagy esetleg valami más leírást tartalmazott. Azokkal a mondatokkal kapcsolatban, amiket ezen utasítás alapján nem gerinccel foglalkozónak jelölt, további információnk nincsen a entitás-címkézés és kapcsolat-kinyerés alfeladatok tekintetében, ezt a kiértékelés során figyelembe is vettük. A gerinccel kapcsolatos mondatok a következő lépésben egy szentiment-kiértékelésen mentek át, amelyeknek során a modell eldöntötte, hogy az adott mondat tartalmaz-e degeneratív elváltozásokat. A nem degeneratív címkével ellátott mondatokat a prompt alapján szintén nem kellett tovább elemezni. Az így

redukált mondathalmazon a modell elvégzi az entitás-címkézés és kapcsolatkinyerés alfeladatokként megfogalmazott műveleteket, amik során kinyeri az *Eváltozásokat* és *Testrészeket*, specifikusan nem írja ki a *Tagadásokat* és a velük kapcsolatban álló *Elváltozásokat*. A negyedik és ötödik lépésben a meghatározott elváltozás-testrész párokat rendezi hármas csoportokba rendezi, hogy megkapja a kívánt level, anatomy, degeneration formát, amit egy szint hozzárendelésével ér el, ilyen szint lehet mondjuk egy specifikus csigolya/porckorong vagy esetleg egy teljes gerincszakasz megnevezése.

A leletek kiértékeléséhez a Azure OpenAI API-t használtuk, a **gpt-4** modell verzióval. Az előfeldolgozás lépésben a promptot úgy állítottuk össze, hogy az előre megírt utasítássorozatot leíró promptot kiegészítettük az eredeti szöveg *findings* mezőjével és a rendszernek szintén megmondtuk, hogy milyen személyiséget "magára öltve" értelmezze azt. A GPT minden esetben rendelkezik ilyen pre-prompttal, amikor elindul egy beszélgetés. A mi esetükben ezt az alapértelmezett pre-promptot felülírtuk azzal az utasítással, hogy: "*As a radiologist, your task involves analyzing radiology reports.*", azaz magyarul "*Mint radiológus, a feladatod az, hogy radiológiai leleteket elemezz ki.*". Ezt a két üzenetet, egyben elküldtük a modellnek (minden lelet esetén), ami a válaszában visszaküldte az eredményt, amit utána utófeldolgozásnak vettettünk alá, ami a kívánt formátumra alakította a kimenetet (lásd 3.2).

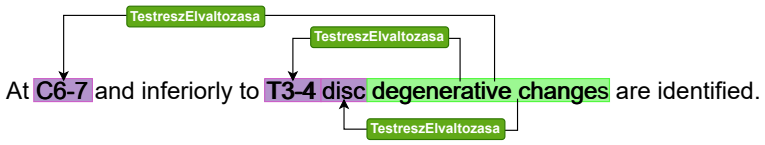
3.4. Kiértékelési szempontok

A kiértékelést egy előre meghatározott szempontrendszer alapján végeztük. A kiértékelés során különböző, a módszerek teljesítmény-mérésére alkalmas mérőszámokat számoltunk a főbb alfeladatokon és szinteken.

3.4.1. Mérőszámok

A módszerek teljesítményének méréséhez a következő mérőszámokat használtuk:

- *Precizitás* (precision) mérésével nézzük, hogy az osztályozott tokenek, relációk közül mekkora hányad releváns.



6. ábra. Példa a reláció-kinyerési feladatra, egy példamondaton belül

- *Fedés* (recall) kiszámításával kaptuk meg, hogy az összesen megtalálandó *Testresz*, *Elvaltozas*, *TestreszElvaltozas* címkékből mennyit találtak meg a modellek.
- *Pontosság* (accuracy) mondta meg, hogy az osztályozások mekkora hányada kapta meg a neki megfelelő címkét.

Továbbá használtuk a fenti *precizitás* és *fedés* mutatókból számolható harmonikus átlagot. Ezt a számot F1-mértéknek (F1-score) hívják. Leggyakrabban az F_1-*score*-t használják a modellek teljesítményének vizsgálatához.

A fentieken felül bevezettünk még egy olyan mérőszámot is, ami azt mutatja meg, hogy az adott módszer átlagosan a leletek mekkora hányadát dolgozta fel. Ez a szám főleg a GPT esetén releváns, mivel a saját megoldásunkról tudjuk, hogy determinisztikusan mindent feldolgoz, amit bemenetként kap, a másik megoldásnál viszont valószínűleg a GPT tokenszám korlátja miatt ez nem ilyen egyértelmű a tapasztalataink alapján.

Mértük továbbá a két módszer válaszidejét is a bemeneti tokenek számának viszonyában. Ezekből az eredményekből a fennálló szerver-különbségek miatt arányt számoltunk, hogy a számadatok megfelelően reprezentálják a két modell inferencia során nyújtott teljesítményét.

3.4.2. Szintek

A kiértékelés során két fő szintet vettünk figyelembe, ezek a szintek külön-külön lettek ellenőrizve az első szint kimenete kihatással volt a másik eredményeire. Mindkét módszer esetén a token osztályozással kezdtük az eredmények kiértékelését, majd utána a reláció-kinyeréssel folytattuk. Amennyiben a token-osztályozásnál valamelyik módszer félrecímkézett esetleg egy *Testrészt*, *Elváltozást* vagy *Tagadást*, akkor ugyanazon a leleten a reláció-kinyerés szinten

már nem tekintettük hibának, ha nem nyert ki olyan összefüggést, aminek az egyik résztvevője az elrontott token lett volna, hiszen a kapott címkézett adatok felett képes csak osztályozni a kapcsolat kinyerő modell.

- **Token-címkézésnél** (példa a 4. ábrán) a kiértékelés során az egyedi BERT kimenet formátuma megegyezett az alapul szolgáló orvosi annotáció formátumával, így az összehasonlítás módszere adta magát. Amennyiben a *Testresz*, *Elvaltozas* entitások[3] megtalálhatóak voltak mindkettőben azok lettek a valódi pozitívok (TP), amik csak az annotációban szerepeltek, azok lettek a hamis negatívok (FN) és, amik csak a modell kimenetében voltak megtalálhatóak, azokat tekintettük hamis pozitívoknak (FP). A hibákat egy táblázatba vezettük, ezeket a mellékelt GitHub repozitóriumban lehet megtekinteni. A GPT kimenetnél külön figyelni kellett még olyan esetekre, ahol a modell időnként szétvágta az eredeti szövegben található entitásokat kisebb részekre (lásd a 3. táblázat), amiket nem tekintettünk hibának, viszont az ilyen esetek megakadályozták egy jól működő automatizált kiértékelő megvalósítását.

- **Reláció-kinyerés** (példa a 6. ábrán). Ennek a lépésnek a kiértékelése a BERT-et használó módszer esetén hasonló volt, mint a *token-osztályzás* lépésben, egy extra előfeldolgozási lépést vett még igénybe, amiben a csővezeték kimenetében szereplő entitások közötti *TestreszResze* és *ElvaltozasResze* relációkat oldottuk fel. Erre a lépésre azért volt szükség, mivel az orvosi annotáció minden *TestreszResze* relációt tartalmazott a résztvevő entitások között, viszont a modell kimenete a legtöbb esetben tartalmazott egy *TestreszElvaltozasa* relációt és több *TestreszResze*/*ElvaltozasResze* relációt, ezeket láncolva lehetett megkapni azokat a párosításokat, amiket utána össze lehetett hasonlítani az orvosi annotációval. A GPT kimenet esetén ennek a részfeladatnak a kiértékelése egyszerűbbnek bizonyult, mivel a modell csoportosítva adta vissza az általa megtalált relációkat, amik tartalmazták az esetenkénti több résztvevőt is, melyekhez a BERT esetén külön lépést kellett

[3]Entitás: Ez alatt értjük az összes *Testresz* és *Elvaltozas* címkével ellátott tokent. Az orvosi annotációban és a kimenetben ezek egy sorszámmal és a token, szövegben található, kezdő- és végpozíciójával vannak kiegészítve.

tenni, hogy feloldjuk a láncokat. A GPT esetén a prompt csoportokat kért a modelltől ezért ezt a feldolgozási lépést nem kellett megtenni a kimenetén.

3. táblázat. Példa az entitások orvos és GPT által történő szétvágására egy rövid leletrészleten

...bony encroachment of the lower lumbar spine...	
radiológus	*GPT*
T19 Testresz 285 301 neural foraminal	level: L4, L5
T20 Elvaltozas 302 314 encroachment	anatomy: neural foramen
T21 Testresz 328 334 lumbar	degeneration: bony encroachment
T22 Testresz 335 340 spine	

4. Kísérletek

A kísérletek során a korábban a módszerek által nem látott leletekből álló teszthalmazon lefuttatuk mindkét módszert, majd az így keletkezett kimeneteket az orvos által készített annotációkkal összehasonlítva kiértékeltük. A kiértékelésekhez főbb alfeladatonként külön fájlhalmaz lett létrehozva, amik segítették a kiértékelési munkát.

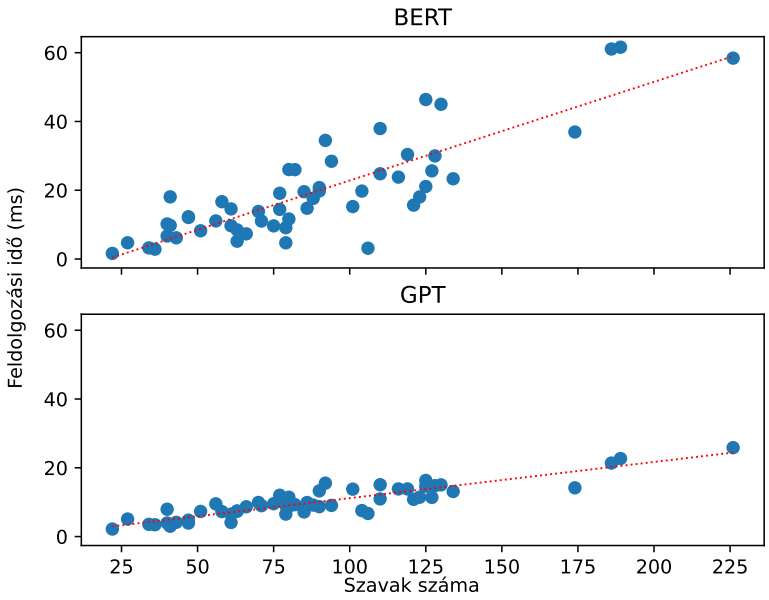
4.1. Eredmények

A kiértékelés során készült táblázatok alapján kiszámoltuk a 3.4.1 alfejezetben kifejtett mérőszámokat, amiket a 4. táblázatban láthatunk. A feldolgozás

4. táblázat. Kiértékelési eredmények

		Pontosság (accuracy) (%)	Precizitás (precision) (%)	Fedés (recall) (%)	F1 (%)
Entitás-osztályozás	GPT	95,78	96,79	87,61	91,97
	BERT	99,16	98,74	97,37	98,05
Reláció-kinyerés	GPT	100,00	100,00	99,62	99.81
	BERT	95,50	94,33	92,26	93,28

kimenetéből szintén kiszámoltuk a mondatfeldolgozási arányt is, erre a GPT esetében volt szükség, mert egyes mondatokat a promptolás ellenére is figyelmen kívül hagyott rendszeresen (megfigyelhetően a leletek végét vágva le), míg a BERT nem hagyott ki mondatokat. A mondatfeldolgozási arány a GPT esetében $95,42\%$ volt a teszhalmazon, míg a BERT az összes mondatot feldolgozta.

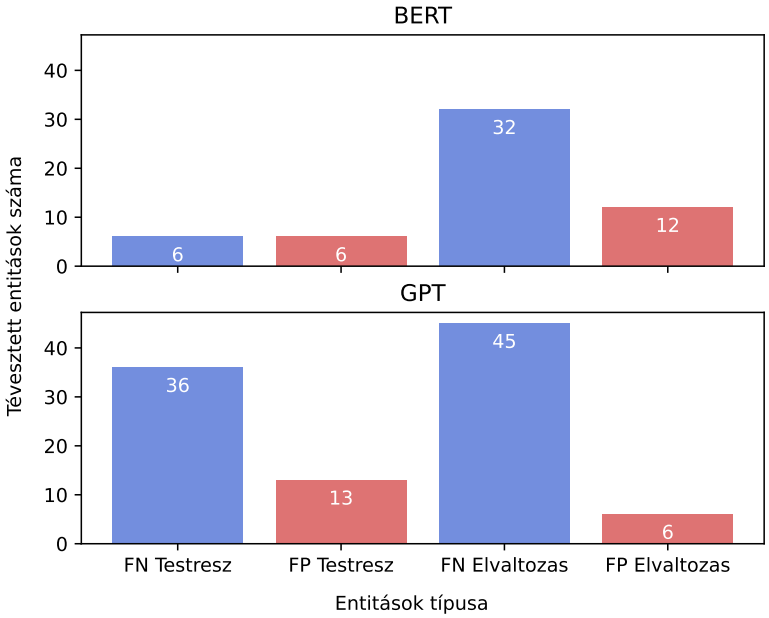


7. ábra. Inferencia idők a szavak számának függvényében

A 7. ábra alapján megállapítottuk[4], hogy a feldolgozási idő a vártaknak megfelelően, arányos a bemeneti szavak számával és megközelítőleg követi a lineáris trend vonalat. Az ábra átláthatósága érdekében egy szélsőséges lelet adatpontját eltávolítottuk, a szöveg a többi vizsgált szöveg hosszához képest is hosszúnak bizonyult és a BERT-et használó csővezeték megközelítőleg *1,5* percig, míg a GPT-4 *40* másodpercig dolgozta fel.

A két módszer két különféle szerveren futott, a BERT-re épülő csővezetéket egy virtuális gépen, dockerizált környezetben futattuk, egy 20 magos AMD

[4]A mérések eredményei megtalálhatóak az elektronikusan csatolt GitHub repozitóriumban: `https://github.com/Yndiliadrin/tdk_adatkinyeresi_modszerek_osszehasonlitasa`

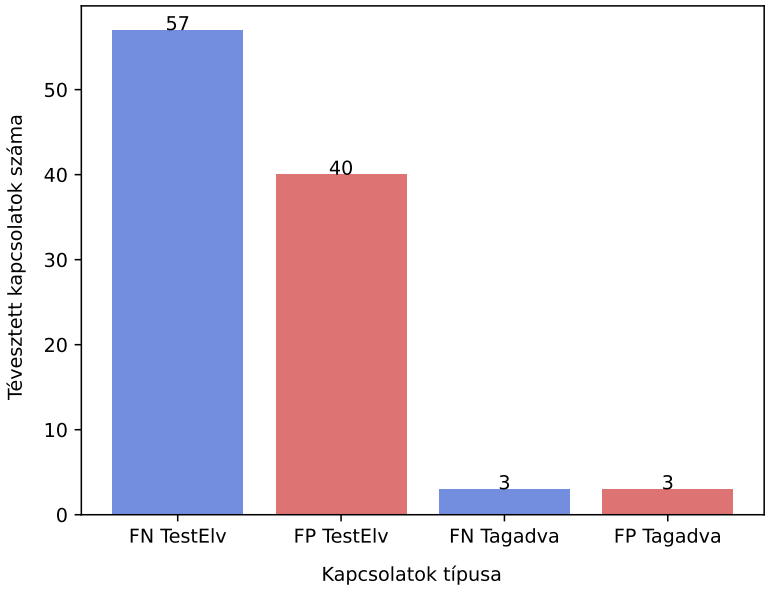


8. ábra. Entitás tévesztések modell szerint

Opteron 6380 (2,5 GHz) proceszoron, 31 GB RAM-mal. A GPT szerevereiről nem sikerült információt szereznünk, azon kívül, hogy valószínűsíthetjük, hogy cél-hardvert használnak a szervereikben, mint például az NVIDIA A100 GPU-k, amikről biztosan tudjuk, hogy a tanítási folyamatban felhasználásra kerültek, valamint valószínűsíthetjük, hogy ezeket a cél hardvert használó gépeket klaszterizálva használják terhelés-elosztás céljából.

4.2. Eredmények értelmezése

Az elért eredmények és a 8. ábra alapján megállapítottuk, hogy az entitás-címkéző alfeladaton a BERT jobb eredményeket ért el, a GPT-s módszerrel szemben az F1-mértékben $6,08\%$-al volt jobb. Ezen a szinten a BERT ***654*** elváltozást és ***808*** testrészt ismert fel, amelyek közül összesen ***38*** volt FN és ***18*** FP. Ezekkel a számokkal szemben a GPT ***243*** elváltozást és ***430*** testrészt ismert fel, amik közül ***81*** volt FN és ***19*** FP. Ezek a számok abban az esetben nem tartalmazzák azokat az eseteket, ahol a modellek szétbontottak, vagy esetleg összevontak elváltozásokat és testrészeket, amik az orvosi annotációban



9. ábra. Az R-BERT-re alapuló kapcsolat-kinyerés tévesztései

esetleg egyben vagy külön voltak megjelölve, amikor ez humán megítélés alapján helyesen tették.

A két módszer pontossága a két fő alfeladaton különbözik, az entitás-osztályzás során a GPT-4 *95,74*%-ot a BERT-et használó módszer (lásd 9. ábra) *99,16*%-ot ért el ezen a mutatón. A reláció-kinyerés során ennek fordítottját figyelhetjük meg, a GPT-4 *100,00*%-ot[5] ért el, amíg a BERT-re alapuló feldolgozó *95,50*%-ot.

A GPT-t használó szövegfeldolgozó jelentősen jobb eredményt ért el a reláció-kinyerés alfeladaton, $6,51$%-os F1-mérték különbséggel lett jobb, mint az egyéni, BERT-re alapuló megoldás. Ezen alfeladaton a GPT ***243*** *TestreszElvaltozasa* és ***21*** *Tagadva* kapcsolatot nyert ki a formázatlan szövegből, amik közül ***1*** FN és ***0*** FP volt. Ugyanezen a feladaton a saját R-BERT-es megoldást ***671*** *TestreszElvaltozasa* és ***147*** *Tagadva* relációt ismert fel, amik közül ***60*** bizonyult FN-nek és ***43*** FP-nek.

[5]Az elektronikusan mellékelt repozitóriumban található *GPT_relations_eval.xlsx* fájl tartalmazza ezt a kiértékelést, láható benne, hogy a GPT-4 az összesen megtalált *264* relációból mindössze **1** FN-t tévesztett.

Megfigyeltük, hogy mindkét módszer kimagaslóan teljesített a tagadások felismerésében, az R-BERT-et használó csővezeték ***147*** *Tagadva* relációt ismert fel, amikből összesen ***3*** FN és ***3*** FP volt, míg a GPT alapú eljárás ***21*** *Tagadva* relációjában ***1*** FN volt. A GPT által megtalált tagadás kapcsolatok száma azért lényegesen kisebb, mert a kiértékelés során csak az általa degeneratívnak ítélt mondatokat néztük, biztonsággal állíthatjuk, hogy a nem degeneratív, valamint a genincen kívüli mondatokban található összes tagadást is felismerte, ám azok nincsenek részeltesen prezentálva a kimeneteben.

Két lehetséges válasz van, hogy miért lehet a tagadások hozzárendelése ilyen pontos. Az első, hogy maga a probléma valószínűleg könyebben tanulható probléma és nem kimondottan domain specifikus, lehetséges, hogy nagy mennyiségű példát láttak a modellek már az előtanításuk során is tagadásokra és azok értelmére. A másik lehetséges magyarázat az, hogy a teszthalmaz alapvetően kevés tagadást tartalmazott, számosítva összesen ***147***-et, ami egy enyhe részrehajlást gyakorolhatott a méréseink eredményeire.

A GPT eredményein javíthatott volna valószínűleg az, ha nincsen az a szokása, hogy hosszabb leletek esetén elvágja a szöveget, esetenként egy mondat közepén is megteszi.

Megállapítottuk továbbá, hogy a BERT magasabb találati számát okozta az is, hogy a GPT által *gerincen kívüli*, *nem degeneratív elváltozást leíró* vagy *degeneratív elváltozást leíró* címkével ellátot mondatok közül, a kimenet csak a *degeneratív elváltozást leíró* mondatokhoz fejtette ki a címkézett entitásokat és kapcsolatokat, így azokat tudtuk kiértékelni a szempontrendszerünk alapján. Azokat az entitásokat és relációkat, amiket az orvosi annotáció a *gerincen kívüli* vagy *nem degeneratív elváltozást leíró* mondatokban jelzett nem vettük figyelembe, a módszer nem kapott semmilyen penalizálást, azért, mert a kimenet nem tartalmazta őket.

4.2.1. Módszerek felhasználhatósága

A GPT-4 felhasználhatóságát nagyban korlátozza a GPT modell természetéből adódó véletlen faktor, nem feltétlenül determinisztikus, hogy egy adott

szövegre milyen eredményt fog kiadni, ez a tulajdonsága API használatával a temperature paraméteren keresztül szabályozható. További akadályt jelenthet, hogy mivel felhőszolgáltatás alapú, a felhaszálók adata nincsen akkora bizalmassággal kezelve, mint egy lokálisan, on-site futtatott megoldásnál. A GPT-4 szöveges orvosi leletek feldolgozásának területén történő alkalmazását tovább akadályozza az a tény is, hogy az Európai Uniós szabályozások tiltják a bizalmas orvosi adatok kiadását, a valódi orvosi leletekkel való munkához etikai engedélyre van szükség, és kórházi együttműködésre. Habár jelenlegi kutatásunkban a kísérletekhez valódi leleteket használtunk, ezek két nyílt lelethalmazból kerültek ki, ahol a leletek számossága és változatossága igen korlátozott. A leletek hozzáférhetősége eleve nagy problémát jelent a kutatók számára, amely megnehezíti mind a mesterséges intelligencián alapuló módszerek fejlesztését, mind azok objektív kiértékelését. A páciens saját belátása szerint, habár ez egyes esetekben szintén tiltott a lelet tulajdonjogát figyelembe véve, mégis fel tudja tölteni a modellnek a személyes adatait. Azonban semmi nem garantálja, hogy a végeredménye az átlag ember számára értelmezhető lesz, valamint a hallucináció és félreértések miatt szélsőséges esetben a saját orvosi ellátását is akadályozhatja ez, amennyiben a humán szakemberek megkerülésére használja fel a páciens.

A GPT-4 további hátrányai közé tartozik az is, hogy mivel általános célú LLM, nem biztos, hogy a további fejlesztéseiben az eszköznek jelentősen javulni fog az orvosi szövegfeldolgozó képessége, hiszen az orvosi leletek továbbra is korlátozottan elérhetők, és a leletek értelmezése nem prioritása a modellnek. Ezért akár véletlenszerű romlásra is képes lehet a megadott feladatokon vagy speciális helyzetekben a későbbi iterációkban, amelyre a felhasználóknak nincs pontos rálátása.

A BERT-et használó megoldás hátránya, hogy egy megfelelő felület mögé kell elhelyezni, önmagában csak nagy mennyiségű, a felhasználó számára értelmezhetetlen szöveget ad vissza az adott inputra, míg a GPT ezzel szemben emberi megfogalmazásba is könnyen át tudja ültetni a kinyert adatokat. Továbbá ez a megoldás jelenleg nem elérhető felhőszolgáltatói oldalon. A felhős kiszolgálás természetesen opció ebben az esetben is, ám ez a GPT jogi hátrányát is magával hozná.

Amennyiben viszont úgy határozunk, hogy on-site futtatjuk a BERT megoldást és építünk köré egy felhasználói alkalmazást, akkor a szerverek és rendszer fenntartásának költségei is minket terhelnének, a szükséges engedélyek megszerzésének problémáján felül.

Egy olyan mesterséges inteligenciára épülő rendszer, ami valós orvosi adatokkal dolgozna, az új Európai Uniós szabályozások szerint a magas kockázatú, azaz *High risk* kategóriába esik. Az ebbe a kategóriába eső mesterséges inteligencia alapú rendszereknek szigorú szűrésen, tesztelésen kell, hogy átmenjenek a valós, nem csak kutatási felhasználáshoz. Ezt egy on-premise megoldással valamivel könnyebben ki lehet vitelezni, mint egy nemzetközi nagyvállalat által szolgáltatott felhő-alkalmazás esetében.

5. Összefoglalás

Kutatásunk közben több szemszögből, tüzetesen megvizsgáltuk a két, nagyban eltérő megoldást, amik közül mindkettő ugyanazt a felmerülő problémát célozta meg. Eredményeinkből kiderült, hogy a két módszer erősségei és hátrányai ellenére megközelítőleg hasonlóan teljesített egy valódi leletekből álló, korábban nem látott teszthalmazon.

Megmutattuk, hogy a GPT-4 önmagában is alkalmas a leletek értelmezésére valódi gerinc-leleteken, a megfelelő pre-prompt/prompt kombináció használatával egy a valóságot jól magyarázó eredményt képes előállítani.

A GPT-4-el ellentétben az általunk fejlesztett, BERT és R-BERT modellekre épülő megoldás szolgáltatásként nem elérhető, ám az on-premise jellegének köszönhetően biztonságosabban és szabályszerűen felhasználató, mint egy specifikusan az orvosi domainre finomhangolt megoldás.

5.1. Továbblépési lehetőségek

A modellek pontosságának növelése egyrészről triviálisnak tekinthető, másrészt további mélyreható kutatást igényelhet. GPT-4 esetén nincsen befolyásunk arra, hogy milyen mennyiségű, minőségű és típusú adaton tanítják a fejlesztői a modellt, így ebben az esetben az újkeletű prompt engineering területe

nyújthat megoldást és biztosíthat jobb eredményeket. A GPT-alapú modellből gyakran jelennek meg új verziók, amelyeket jelen dolgozat nem vizsgált, ám elképzelhető, hogy a lelet-strukturálási képességük is változik ezzel, a későbbi modell-változatoknál ezért igen indokolt lehet újabb kutatást végezni a BERT és GPT közötti összehasonlításokról. A BERT-alapú megoldásunk esetén a még több jó minőségű adat használata a tanítások során természetesen egy nagyon hatékony mód a módszer pontosságának növelésére.

Egy GPT-4-et magába integráló leletfeldolgozó rendszer megvalósítása jelen állás szerint számos jogi aggályt vonna maga után. Maga az LLM-ekre épülő, leletezést felgyorsító, betegedukációt elősegítő alkalmazás készítése viszont egy olyan ötlet, amit már a világ több pontján is kutatnak, és fejlesztéseket tesznek a cél elérése érdekében. Hazánkban például a Budai Egészségközpont és egyetemünk által fejlesztett iLelet projekt keretében magyar nyelvű radiológiai leletek feldolgozását és vizualizáltan történő megjelenítését céloztuk meg, amelynek fejlesztésében én is részt vettem.

6. Köszönetnyílvánítás

A kutatás során többen segítették munkámat, nekik szeretném külön megköszönni munkájukat. Köszönöm Szabó Ledenyi Klaudia doktorandusznak a szakmai segítségét és a kiértékelési szabályrendszer megalkotásában nyújtott tanácsait. Kicsi Andrásnak, a Szegedi Tudományegyetem adjunktusának szintén köszönöm a szabályrendszer megalkotásában nyújtott értékes segítségét, valamint jelen dolgozat lektorálását. Továbbá szeretném megköszönni Balogh Andrásnak dolgozatom nyelvhelyességi ellenőrzését.

A kutatás az Európai Unió támogatásával valósult meg, az RF-2.3.1.1-21-2022-00004 azonosítójú, Mesterséges Intelligencia Nemzeti Laboratórium projekt keretében. A kutatás a TKP2021-NVA-09 projekt támogatásával készült. A TKP2021-NVA-09 számú projekt Magyarország Innovációs és Technológiai Minisztériumának támogatásával valósult meg a Nemzeti Kutatási, Fejlesztési és Innovációs Alapból, a TKP2021-NVA támogatási finanszírozási keret alapján.

7. Melléklet

A dolgozat elkészítéséhez használt adatok, a kiértékelésük során keletkezett táblázatok a tdk_adatkinyeresi_modszerek_osszehasonlitasa GitHub repozitóriumban[6] érhetőek el.

A reláció-kinyeréshez használt R-BERT alap implementáció a következő címen érhető el: `https://github.com/monologg/R-BERT?tab=readme-ov-file`

[6] `https://github.com/Yndiliadrin/tdk_adatkinyeresi_modszerek_osszehasonlitasa`

Hivatkozások

[1] *ChatGPT*. URL: https://chatgpt.com/.

[2] *Gemini*. URL: https://gemini.google.com/app (elérés dátuma 2024. 11. 18.).

[3] *AUTORAD*. URL: https://medicalonline.hu/gyogyitas/cikk/autorad.

[4] Klaudia Szabó Ledenyi; András Kicsi; László Vidács. *A Deep Dive into GPT-4's Data Mining Capabilities for Free-Text Spine Radiology Reports - DATA 2024*. 2024. URL: https://www.insticc.org/node/TechnicalProgram/data/2024/presentationDetails/127651.

[5] Alec Radford és Karthik Narasimhan. „Improving Language Understanding by Generative Pre-Training". (2018). URL: https://api.semanticscholar.org/CorpusID:49313245.

[6] Jacob Devlin és tsai. „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* 1 (2018. okt.), 4171–4186. old. URL: https://arxiv.org/abs/1810.04805v2.

[7] Alistair E.W. Johnson és tsai. „MIMIC-III, a freely accessible critical care database". *Scientific data* 3 (2016. máj.). ISSN: 2052-4463. DOI: 10.1038/SDATA.2016.35. URL: https://pubmed.ncbi.nlm.nih.gov/27219127/.

[8] *Transcribed Medical Transcription Sample Reports and Examples - MTSamples*. URL: https://mtsamples.com/ (elérés dátuma 2024. 11. 18.).

[9] Jinhyuk Lee és tsai. „BioBERT: a pre-trained biomedical language representation model for biomedical text mining". *Bioinformatics* 36.4 (2019. jan.), 1234–1240. old. DOI: 10.1093/bioinformatics/btz682. URL: http://arxiv.org/abs/1901.08746%20http://dx.doi.org/10.1093/bioinformatics/btz682.

[10] *BookCorpus Dataset | Papers With Code*. URL: https://paperswithcode.com/dataset/bookcorpus (elérés dátuma 2024. 11. 18.).

[11] *WebText Dataset | Papers With Code*. URL: https://paperswithcode.com/dataset/webtext (elérés dátuma 2024. 11. 18.).

[12] *Index of /enwiki/*. URL: https://dumps.wikimedia.org/enwiki/ (elérés dátuma 2024. 11. 18.).

[13] Renqian Luo és tsai. „BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining". *Briefings in Bioinformatics* 23.6 (2022. okt.). ISSN: 14774054. DOI: 10.1093/bib/bbac409. URL: https://arxiv.org/abs/2210.10341v3.

[14] Iz Beltagy, Kyle Lo és Arman Cohan. „SciBERT: A Pretrained Language Model for Scientific Text". *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference* (2019. márc.), 3615–3620. old. DOI: 10.18653/v1/d19-1371. URL: https://arxiv.org/abs/1903.10676v3.

[15] Xieling Chen és tsai. „A bibliometric analysis of natural language processing in medical research". *BMC Medical Informatics and Decision Making* 18.1 (2018. márc.), 1–14. old. ISSN: 14726947. DOI: 10.1186/S12911-018-0594-X/TABLES/10. URL: https://link.springer.com/articles/10.1186/s12911-018-0594-x%20https://link.springer.com/article/10.1186/s12911-018-0594-x.

[16] Matthew C. Chen és tsai. „Deep Learning to Classify Radiology Free-Text Reports". *https://doi.org/10.1148/radiol.2017171115* 286.3 (2017. nov.), 845–852. old. ISSN: 15271315. DOI: 10.1148/RADIOL.2017171115. URL: https://pubs.rsna.org/doi/10.1148/radiol.2017171115.

[17] Bernal Jiménez Gutiérrez és tsai. „Thinking about GPT-3 In-Context Learning for Biomedical IE? Think Again". *Findings of the Association for Computational Linguistics: EMNLP 2022* (2022. márc.), 4526–4541. old. DOI: 10.18653/v1/2022.findings-emnlp.329. URL: https://arxiv.org/abs/2203.08410v3.

[18] Yinhan Liu és tsai. „RoBERTa: A Robustly Optimized BERT Pretraining Approach". *arXiv* (2019. júl.). URL: https://arxiv.org/abs/1907.11692v1.

[19] Bo Guo, Huaming Liu és Lei Niu. „Integration of natural and deep artificial cognitive models in medical images: BERT-based NER and relation extraction for electronic medical records". *Frontiers in Neuroscience* 17 (2023. szept.), 1266771. old. ISSN: 1662453X. DOI: 10.3389/FNINS.2023.1266771/BIBTEX.

[20] Matjaž Gams és tsai. „Developing a Medical Chatbot: Integrating Medical Knowledge into GPT for Healthcare Applications". *Ambient Intelligence and Smart Environments* (2024. jún.). DOI: 10.3233/AISE240018. URL: https://www.researchgate.net/publication/381510388.

[21] James C.L. Chow, Leslie Sanders és Kay Li. „Impact of ChatGPT on medical chatbots as a disruptive technology". *Frontiers in Artificial Intelligence* 6 (2023. ápr.), 1166014. old. ISSN: 26248212. DOI: 10.3389/FRAI.2023.1166014/BIBTEX.

[22] Chiranjib Chakraborty és tsai. „Overview of Chatbots with special emphasis on artificial intelligence-enabled ChatGPT in medical science". *Frontiers in Artificial Intelligence* 6 (2023. okt.), 1237704. old. ISSN: 26248212. DOI: 10.3389/FRAI.2023.1237704/BIBTEX.

[23] Arun James Thirunavukarasu és tsai. „Large language models in medicine". *Nature Medicine 2023* 29.8 (2023. júl.), 1930–1940. old. ISSN: 1546-170X. DOI: 10.1038/s41591-023-02448-8. URL: https://www.nature.com/articles/s41591-023-02448-8.

[24] Madhumita Sushil, Simon Šuster és Walter Daelemans. „Are we there yet? Exploring clinical domain knowledge of BERT models". *Proceedings of the 20th Workshop on Biomedical Language Processing, BioNLP 2021* (2021), 41–53. old. DOI: 10.18653/V1/2021.BIONLP-1.5. URL: https://aclanthology.org/2021.bionlp-1.5.

[25] Shanchan Wu és Yifan He. „Enriching Pre-trained Language Model with Entity Information for Relation Classification”. *International Conference on Information and Knowledge Management, Proceedings* (2019. máj.), 2361–2364. old. DOI: 10.1145/3357384.3358119. URL: https://arxiv.org/abs/1905.08284v1.

[26] *Brat rapid annotation tool*. URL: https://brat.nlplab.org/ (elérés dátuma 2024. 11. 18.).

[27] Andras Kicsi és tsai. „Automatic Classification and Entity Relation Detection in Hungarian Spinal MRI Reports”. *Proceedings - 2021 IEEE/ACM 3rd International Workshop on Software Engineering for Healthcare, SEH 2021* (2021. jún.), 13–19. old. DOI: 10.1109/SEH52539.2021.00010.

[28] R. Ramachandran és K. Arutchelvan. „Named entity recognition on biomedical literature documents using hybrid based approach”. *Journal of Ambient Intelligence and Humanized Computing* (2021. márc.), 1–10. old. ISSN: 18685145. DOI: 10.1007/S12652-021-03078-Z/FIGURES/5. URL: https://link.springer.com/article/10.1007/s12652-021-03078-z.

[29] Pengfei Liu és tsai. „Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing”. *ACM Computing Surveys* 55.9 (2021. júl.). ISSN: 15577341. DOI: 10.1145/3560815. URL: https://arxiv.org/abs/2107.13586v1.