Táblázatos kérdés-megválaszolás magyar nyelven

Tóth Gábor, Programtervező informatikus szak Msc, II. évf

Dr. Farkas Richárd József egyetemi docens, Szegedi Tudományegyetem

# Bevezetés

Az utóbbi években a mesterséges intelligencia és a természetes nyelvfeldolgozás (NLP) területén a nagy nyelvi modellek (LLM-ek) jelentős fejlődésen mentek keresztül. Ezek a modellek, mint például a GPT, a BERT és a T5, képesek nagymennyiségű adat feldolgozására és számos különböző nyelvi feladat végrehajtására, beleértve a szövegértést, fordítást, összegzést és a kérdés-válaszolást is. Az LLM-ek legfőbb előnye, hogy képesek általánosítani és összetett nyelvi struktúrákat értelmezni, ami kiemelkedően fontossá teszi őket a nyelvi adatok feldolgozásában. Bár ezek a modellek elsősorban az angol nyelvre optimalizáltak, egyre nagyobb igény mutatkozik arra, hogy más nyelveken, például magyarul is sikeresen alkalmazhatók legyenek.

Az egyik különösen érdekes alkalmazási terület a táblázatos kérdés-válasz (TQA) feladat, amely során a modellnek strukturált adatokból, táblázatokból kell a feltett kérdésekre választ adnia. A TQA-feladatok jelentősége az adatbázisokban és különböző dokumentumokban elérhető strukturált információ hasznosításában rejlik, amelyeket hatékony kérdés-válaszoló rendszerek segítségével lehet automatizálni. Az angol nyelven számos kutatás foglalkozik TQA-kkal, és több jól ismert TQA-adathalmaz is rendelkezésre áll, amelyek segítségével az algoritmusokat és nyelvi modelleket fejleszteni lehet. Azonban magyar nyelven ezek az erőforrások korlátozottak, és jelenleg nem áll rendelkezésre olyan TQA-adatbázis, amely magyar nyelven készült kérdések és válaszok vizsgálatát tenné lehetővé.

Jelen dolgozat célja egy magyar nyelvű TQA-adathalmaz létrehozása, amely a magyar Wikipédia táblázatos adatait és szöveges tartalmát hasznosítja. A Wikipédia szócikkeiben számos táblázatos formában közölt információ található, amelyek jól strukturált és változatos adatokat tartalmaznak a legkülönfélébb témakörökben. Az ezekből az adatokból készített TQA-adathalmaz lehetőséget nyújt a magyar nyelvű LLM-ek tesztelésére, valamint a TQA-algoritmusok hatékonyságának vizsgálatára. A kutatás során egy automatizált adatgyűjtési és -feldolgozási eljárást alakítottam ki, amely a Wikipédia táblázatait és a kapcsolódó szöveges tartalmakat kérdés-válasz párokba rendezi, így egy jól strukturált adatbázis jön létre, amely alkalmas a TQA-modellek kiértékelésére.

A dolgozat további célja, hogy különböző nagy nyelvi modellek teljesítményét és megbízhatóságát mérje fel a magyar nyelvű TQA-adathalmazon, összehasonlítva a különböző algoritmusokat és technikai megközelítéseket. Ennek érdekében többféle LLM-et alkalmazok, és különböző TQA-feladatok megoldására használt algoritmusokat tesztelek. Az eredmények alapján összehasonlítom a modellek hatékonyságát és pontosságát, különös tekintettel a magyar nyelv sajátosságaira. A kutatás hozzájárul a magyar nyelvű természetes nyelvfeldolgozási kutatások bővítéséhez, mivel ez az adatbázis elősegíti a magyar nyelvű TQA-rendszerek fejlesztését és kiértékelését, amelyek hosszú távon alapvető fontosságúak lehetnek a magyar nyelvű digitális szolgáltatások és alkalmazások számára.

Összefoglalva, ez a dolgozat a magyar nyelvű TQA-kutatásban jelentős előrelépést jelenthet, hiszen megteremti egy új típusú magyar nyelvű adathalmaz alapját, és lehetőséget biztosít arra, hogy a magyar nyelven is teszteljük a legkorszerűbb LLM-eket.

# LLM

A nagy nyelvi modellek (LLM-ek) a mesterséges intelligencia (AI) és a természetes nyelvfeldolgozás (NLP) területén kifejlesztett komplex rendszerek, amelyek képesek nagy mennyiségű szöveg adatainak értelmezésére, elemzésére és feldolgozására. Az LLM-ek alapját a neurális hálózati modellek képezik, amelyek mérete és bonyolultsága jelentősen meghaladja a korábbi NLP-modellekét. A legmodernebb LLM-ek, mint például a GPT (Generative Pre-trained Transformer) modellek, a BERT (Bidirectional Encoder Representations from Transformers), valamint a T5 (Text-To-Text Transfer Transformer), több milliárd paramétert tartalmaznak, és ennek köszönhetően széles körű feladatokat képesek ellátni, a szövegértelmezéstől kezdve a kérdés-válaszoláson át a szövegkészítésig.

Az LLM-ek jellemzően deep learning technológián alapulnak, és ezeknek az egyik legismertebb megvalósítása a „transformer” architektúra. A transformer-modellt Vaswani és munkatársai 2017-ben vezették be, és rövid időn belül meghatározóvá vált az NLP területén. A transformer alapvetően „self-attention” mechanizmusokat használ, amelyek lehetővé teszik, hogy a modell az input szöveg különböző részei között hosszú távú kapcsolatokat hozzon létre és felismerje azokat az összefüggéseket, amelyek alapvetőek a nyelvi megértéshez.

Az LLM-ek előképzéséhez (pre-training) általában hatalmas mennyiségű szöveganyagot használnak, amelyek tartalmazzák az interneten fellelhető különféle nyilvános adatforrásokat, könyveket és más szövegeket. Az előképzés során a modellek arra tanulnak, hogy szavak, kifejezések és szövegblokkok között mintázatokat és kapcsolódásokat fedezzenek fel, ami képessé teszi őket a nyelv természetes feldolgozására. Az előképzés után jellemzően finomhangolási (fine-tuning) lépéseken mennek keresztül, amelyek során az adott feladatra optimalizálják őket, például kérdés-válaszolásra vagy fordításra.

Az LLM-ek egyik legnagyobb előnye, hogy sokféle nyelvi feladatra használhatók, méretük és komplexitásuk lehetővé teszi, hogy az alapvető szövegértési feladatokon túlmenően is teljesítsenek. Ezek a modellek képesek szöveget generálni, fordítani, kérdésekre válaszolni, és segíteni különféle kreatív írási feladatokban, például történetmesélésben és összegzésben. A GPT-modellsorozat, amelyet az OpenAI fejlesztett, például jól ismert a szöveggenerálási képességeiről. A GPT-3 modell több mint 175 milliárd paramétert tartalmaz, és az egyik legnagyobb nyelvi modell, amely eddig készült.

A BERT-modell, amelyet a Google fejlesztett ki, inkább arra specializálódott, hogy szövegértelmezési feladatokat végezzen el, mint például a szövegkategorizálás és a kérdés-válaszolás. A BERT kétirányú tanulási mechanizmusa lehetővé teszi, hogy a modell mind a bal, mind a jobb kontextust figyelembe vegye az egyes szavak jelentésének értelmezésekor, ami különösen hatékonnyá teszi az NLP-feladatok során.

Annak ellenére, hogy az LLM-ek sikeresen alkalmazhatók az angol nyelvű feladatokban, más nyelvek esetében korlátozottabb eredményeket nyújtanak. Ennek egyik oka, hogy az LLM-ek előképzéséhez szükséges hatalmas mennyiségű adathoz nem minden nyelv esetében férnek hozzá, így például a magyar nyelvre optimalizált modellek kifejlesztése továbbra is kihívást jelent. Az LLM-ek előképzése gyakran az interneten elérhető, angol nyelvű szövegeken történik, és ezáltal egyes nyelvek, különösen a kisebb nyelvek, mint a magyar, alulreprezentáltak maradnak. Az olyan kutatások, amelyek magyar nyelvű kérdés-válaszoló adatbázisokat hoznak létre, alapvetően fontosak a magyar nyelvű LLM-ek fejlesztésében és azok teljesítményének javításában.

# TQA feladat

A Táblázatos Kérdés-Válasz (Table Question Answering, TQA) feladat a természetes nyelvfeldolgozás egyik speciális alkalmazási területe, ahol a modell célja, hogy strukturált adatokból (táblázatokból) válaszoljon a feltett kérdésekre. A TQA különösen hasznos, amikor nagy mennyiségű adatot szeretnénk értelmezni és kinyerni jól meghatározott kérdések megválaszolásával, például pénzügyi táblázatok, sportstatisztikák vagy adatbázisok esetében.

A TQA-feladat során a modell egy vagy több táblázatot kap bemenetként, valamint egy természetes nyelven megfogalmazott kérdést, amelyre a táblázatban található információk alapján kell válaszolnia. A feladat sikeres végrehajtásához a modellnek képesnek kell lennie arra, hogy a kérdést lefordítsa egy logikai vagy matematikai utasításokra bontott formára, majd ez alapján a táblázat megfelelő celláiban található adatokat azonosítsa és felhasználja. Ez magában foglalhat egyszerű adatlekérdezéseket (pl. „Mennyi az X termék ára?”), összetett műveleteket (pl. „Melyik évben volt a legnagyobb növekedés az Y termék értékesítésében?”), valamint több lépésben történő információkombinációkat is.

A TQA-feladat számos technikai kihívással jár:

1. **Természetes nyelvi megértés és táblázatértés integrációja**: A modellnek képesnek kell lennie arra, hogy a kérdésben szereplő nyelvi utasításokat összekapcsolja a táblázat strukturált elemeivel (sorok, oszlopok, cellák).
2. **Szintaktikai és szemantikai összefüggések felismerése**: Gyakran előfordul, hogy a táblázatban található adatot különböző logikai kapcsolatokon vagy szemantikai összefüggéseken keresztül kell értelmezni. Például a „legnagyobb növekedés” vagy „legkisebb érték” kérdések megválaszolásához a modellnek összehasonlító műveleteket kell végrehajtania.
3. **Számítási és logikai műveletek végrehajtása**: A kérdések egy része számításokat igényel, mint például összegzést, átlagolást vagy százalékszámítást. Ehhez a modellnek egyaránt szüksége van matematikai és logikai képességekre.
4. **Adaptálás több nyelvhez**: Mivel a legtöbb TQA-modellt elsősorban angol nyelven fejlesztik és tanítják, a kisebb nyelveken (például magyarul) való alkalmazásuk jelentős kihívást jelent, különösen a helyi nyelvi és táblázati sajátosságok figyelembevétele szempontjából.

A TQA-feladatok megoldására számos különböző algoritmust és modellt használnak. Az egyik legelterjedtebb megközelítés a transformer alapú modellek, mint például a TaBERT (Table-BERT), a TAPAS és a T5, amelyek különösen jól teljesítenek táblázatok feldolgozásában. Ezek a modellek kifejezetten úgy lettek tervezve, hogy képesek legyenek integrálni a táblázat szerkezetét a nyelvi modell kontextusába, így megkönnyítve a táblázatokból történő információkinyerést.

A TQA-feladatokra használt modellek gyakran finomhangolásra kerülnek speciális TQA-adathalmazokon, amelyek tartalmazzák a táblázatokat, a kapcsolódó kérdéseket és a helyes válaszokat. Ilyen adathalmaz például a WikiTableQuestions vagy a SQA (Sequential Question Answering), amelyeket széles körben használnak a TQA-rendszerek tesztelésére és fejlesztésére.

A TQA-feladatok jövőbeli fejlesztési irányai közé tartozik a többnyelvű és nyelvfüggetlen TQA-rendszerek kidolgozása, amelyek lehetővé teszik a modellek számára, hogy különböző nyelveken is jól teljesítsenek, például a magyar Wikipédián található táblázatokra is alkalmazhatók legyenek. Emellett növekvő hangsúlyt kap az a kérdés, hogy a modellek hogyan tudják elkerülni a hibás válaszokat, különösen bonyolult kérdések esetén, amelyek több táblázat vagy összetett műveletek bevonását igénylik.

# Adatbázisok

Számos TQA adatbázis létezik, amelyek különféle típusú kérdés-válasz feladatokat támogatnak és különféle táblázatos struktúrákat tartalmaznak. Ezek az adathalmazok a TQA modellek fejlesztésének és tesztelésének alapját képezik.

1. \*\*WikiTableQuestions\*\* Az egyik legismertebb TQA adatbázis, amely a Wikipedia táblázataiból származik. Az adatbázisban található táblázatokhoz természetes nyelven megfogalmazott kérdések tartoznak, amelyeket a modellnek meg kell válaszolnia. A kérdések gyakran tartalmaznak egyszerű lekérdezéseket (pl. adott érték megtalálása), de vannak bonyolultabb, több lépéses műveleteket is igénylő feladatok is. A WikiTableQuestions az egyik legelterjedtebb TQA adatbázis, különösen a szöveg és táblázatos adat integrációjára építő modellek fejlesztéséhez.

2. \*\*SQA (Sequential Question Answering)\*\* Ez az adatbázis úgy lett kialakítva, hogy egymást követő kérdések sorozatát tartalmazza ugyanazon táblázatokhoz, ahol a kérdések gyakran egymásra épülnek. Az SQA ezért különösen alkalmas olyan TQA modellek fejlesztésére, amelyek több lépésben megfogalmazott kérdéseket is képesek kezelni, mivel gyakran szükséges a korábbi kérdések válaszainak figyelembevétele az újabb kérdések megválaszolásakor.

3. \*\*WikiSQL\*\* A WikiSQL adatbázis SQL-lekérdezéseket tartalmaz természetes nyelven megfogalmazott kérdésekkel, amelyek táblázatos adatokra vonatkoznak. Bár ez az adatbázis inkább SQL-re épít, hasznos a TQA modellek fejlesztésében is, mivel a természetes nyelvi kérdéseket kapcsolja össze a táblázatokkal. Ez az adathalmaz különösen azoknak a modelleknek hasznos, amelyek logikai vagy strukturált lekérdezésként dolgozzák fel a természetes nyelvi inputot.

4. \*\*TabFact\*\* A TabFact adatbázis a táblázatos tényellenőrzésre fókuszál. Az adatbázis olyan természetes nyelvi állításokat tartalmaz, amelyek alapján a modelleknek el kell dönteniük, hogy a megadott táblázat adatai alapján igazak vagy hamisak-e. A TabFact ezért inkább tényellenőrzési feladatokra használatos, de TQA-feladatokban is alkalmazható, amikor az állítások és kérdések értelmezése táblázatos kontextusban történik.

5. \*\*HybridQA\*\* A HybridQA adatbázis kombinált kérdés-válasz feladatokat tartalmaz, amelyek táblázatos és szöveges forrásokat is igényelnek a kérdések megválaszolásához. Itt a válaszok egyaránt követelhetik meg táblázatok és szöveges bekezdések használatát, így az adatbázis különösen alkalmas komplex TQA-feladatok gyakorlására, ahol a modelleknek képesnek kell lenniük többféle információforrást integrálni.

6. \*\*TAT-QA (Table-and-Text Question Answering)\*\* A TAT-QA adatbázis olyan kérdéseket tartalmaz, amelyek szintén táblázatos és szöveges forrásokra egyaránt építenek. Különösen pénzügyi és számítási feladatokat tartalmaz, ahol a modelleknek össze kell vonniuk a szöveges leírásokat és a táblázatok adatait a helyes válasz előállításához. A TAT-QA az üzleti és pénzügyi szektorhoz kapcsolódó TQA-feladatokban használatos, ahol az egyes kérdések több lépésben történő számításokat igényelhetnek.

Ezek az adatbázisok különféle típusú TQA modellek fejlesztését teszik lehetővé, így mind a szöveg-alapú, mind a táblázat-orientált információfeldolgozásban hasznosak, hozzájárulva a TQA-technológia fejlődéséhez.agyar TQA adatbázisok

# Magyar TQA adatbázis

A magyar nyelvű TQA (Táblázatos Kérdés-Válasz) adatbázis létrehozása a magyar Wikipédián található táblázatok és kapcsolódó szövegek automatizált gyűjtésével és feldolgozásával történt. A cél az volt, hogy olyan kérdés-válasz párokat hozzunk létre, amelyek lehetővé teszik a nagyméretű nyelvi modellek tesztelését és finomhangolását a magyar nyelvű táblázatos kérdés-válasz feladatokhoz. Az alábbiakban részletesen bemutatjuk az adatbázis összeállításának folyamatát, a Wikipédiából való adatgyűjtést, az adatok tisztítását és az adathalmaz struktúráját.

**1. Adatgyűjtési Folyamat**

A magyar Wikipédiában található cikkek jelentős része tartalmaz olyan táblázatokat, amelyek különféle adatokat és kategóriákat rendeznek struktúrált formában. Az adatgyűjtési folyamat során célzottan a következő típusú információkat gyűjtöttük:

* **Táblázatos adatok**: A cikkekben található táblázatok, amelyek számos adattípust, például statisztikai adatokat, időpontokat, országokat, földrajzi helyeket, vagy történelmi eseményekre vonatkozó információkat tartalmaznak.
* **Kapcsolódó szöveges kontextus**: A táblázatok körül található szöveges információkat is feldolgoztuk, mivel ezek gyakran magyarázzák vagy részletezik a táblázatok tartalmát. Ez a kontextus hasznos a kérdések megalkotásához, mivel a szöveg utalhat a táblázat egyes celláira, és további jelentést vagy magyarázatot ad az adatokhoz.

Az adatgyűjtés a következő lépésekből állt:

* **Webszkréperek és API-k használata**: A magyar Wikipédiából történő adatkinyeréshez Python-alapú webszkrépereket és a MediaWiki API-t használtuk, amely lehetővé tette a cikkek szövegének, a táblázatok HTML-kódjának és a formázott adatainak automatikus letöltését.
* **Adatstruktúra kivonása**: Az összegyűjtött táblázatok HTML-struktúrájának feldolgozása során az adatokat egy strukturált formába alakítottuk át, amelyben minden táblázatsor és oszlop különálló entitásként kezelhető. Így könnyebben társíthatóak a táblázat celláihoz kapcsolódó kérdések.

**2. Adattisztítás és Feldolgozás**

Az összegyűjtött adatok további feldolgozást igényeltek annak érdekében, hogy pontos, egységes és használható adathalmazt hozzunk létre. A tisztítás során a következő feladatokat végeztük el:

* **Duplikációk eltávolítása**: Mivel a Wikipédiában előfordulhat, hogy egyes táblázatok ismétlődnek vagy hasonló adatokat tartalmaznak, ezért eltávolítottuk a redundáns bejegyzéseket, hogy az adathalmaz minél változatosabb és informatívabb legyen.
* **Hiányzó adatok kezelése**: A táblázatok egyes cellái esetenként hiányosak vagy nem tartalmaznak releváns információt. Ezeket a cellákat vagy kitöltöttük, ha a kontextus alapján lehetett következtetni a helyes adatértékre, vagy jelöltük, hogy ezek az adatok nem használhatóak a kérdés-válasz feladatban.
* **Adatok típus szerinti rendezése**: Az adatokat különböző típusok szerint csoportosítottuk (pl. dátum, szám, szöveg), hogy a későbbi feldolgozás során könnyebben használhatók legyenek. Például az időpontokat és számadatokat formailag és tartalmilag is egységesítettük.

**3. Kérdés-Válasz Párok Generálása**

A kérdés-válasz párok létrehozása során célunk volt, hogy a táblázatokhoz kapcsolódó minél több típusú kérdést lefedjük. A kérdés-válasz párok generálásánál figyelembe vettük, hogy a kérdések:

* **Egyszerű információt kérjenek**: Például egy adott cellában szereplő konkrét adat megkérdezése, mint „Mi Magyarország fővárosa?” egy országokra vonatkozó táblázatban.
* **Összetett lekérdezéseket igényeljenek**: Olyan kérdések, amelyek több cella vagy sor összehasonlítását, összegzését vagy kiszámítását követelik meg, például „Melyik ország népessége a legnagyobb?”.
* **Időbeli és logikai kapcsolatokat kérdezzenek**: Például „Hány év telt el Magyarország uniós csatlakozása óta?” vagy „Melyik esemény történt előbb?” Az ilyen kérdések célja a modell logikai következtetési képességének vizsgálata.

A kérdések generálása automatikus és félig automatikus folyamat volt. Automatikusan generált kérdések esetén szkripteket használtunk, amelyek különféle kérdéstípusokat tudtak létrehozni a táblázatok alapján, míg félig automatikus kérdésgenerálásnál emberi ellenőrzés és finomhangolás is szükséges volt a relevancia biztosítása érdekében.

**4. Adathalmaz Struktúrája és Formátuma**

Az így összeállított adathalmaz egy strukturált formátumban került tárolásra, amely biztosítja, hogy a TQA-feladatokhoz könnyen felhasználható legyen. Az adathalmaz a következő elemeket tartalmazza:

* **Táblázat**: A forrástáblázat maga, amely a Wikipédiáról származik, és strukturált formában van tárolva.
* **Kérdések**: Minden táblázathoz kapcsolódóan több kérdés-válasz pár található. A kérdések különböző nehézségűek és típusúak, hogy különféle modellek képességeit mérhessük.
* **Kapcsolódó szöveges kontextus**: Az adatbázisban szerepel a táblázatok környezetében lévő szöveges tartalom is, amely további információt nyújthat a válaszadás során.
* **Válaszok**: Minden kérdéshez tartozik egy referencia-válasz, amely az elvárt kimenetként szolgál a kiértékelés során.

**5. Formátum és Tárolás**

A magyar TQA-adatbázis többféle formátumban került tárolásra, hogy a különböző elemzési módszerekhez és TQA-megoldásokhoz könnyen hozzáférhető legyen:

* **CSV vagy TSV fájlok**: Az egyszerűbb feldolgozás érdekében minden táblázat, kérdés és válasz CSV vagy TSV fájlok formájában került tárolásra, amely formátumok könnyen beolvashatók adatfeldolgozó szkriptekbe.
* **JSON formátum**: A hierarchikus adatstruktúrákat jobban támogató JSON formátum szintén elérhető, ahol a táblázat, a kérdések és a válaszok egy objektumként vannak tárolva. Ez a formátum lehetővé teszi, hogy a kérdések és válaszok közvetlenül a táblázathoz kapcsolódjanak, így jobban illeszkedik a modern NLP (Natural Language Processing) modellekhez.
* **SQL-adatbázis**: A strukturált relációs lekérdezések érdekében SQL-adatbázisban is tároltuk az adatokat. Ez lehetővé teszi a különböző szűrések és lekérdezések futtatását a táblázatok és kérdések alapján, ami hasznos a kísérletekhez.

**6. Adathalmaz Használhatósága és Potenciálja**

Az elkészült magyar TQA-adatbázis széleskörűen felhasználható a magyar nyelvű nagy nyelvi modellek (LLM-ek) képességeinek értékelésére, különös tekintettel a táblázatos kérdés-válasz feladatokra. Az adatbázis alkalmas különféle modellek tesztelésére és finomhangolására, és hozzájárulhat a magyar nyelvű NLP fejlődéséhez. Az adatbázis egyedülálló abban, hogy kifejezetten a magyar Wikipédia tartalmára épül, amely hiteles, magyar nyelvű adatokat és szövegeket biztosít a TQA modellek számára.

# Felhasznált LLM lista

A kutatásban felhasznált LLM-ek listája a következő modelleket tartalmazza, amelyek a magyar nyelvű táblázatos kérdés-válasz feladatokhoz nyújtanak alapot. Ezek a modellek különböző architektúrákkal és előzetes tréningfolyamatokkal rendelkeznek, amelyek révén hatékonyan képesek a természetes nyelv feldolgozására és a táblázatokban rejlő információk kinyerésére.

1. \*\*BERT (Bidirectional Encoder Representations from Transformers)\*\*

A BERT egy kétirányú transformer-alapú modell, amelyet különféle nyelvfeldolgozási feladatokban széles körben használnak. Bár eredetileg nem táblázatokhoz fejlesztették, a BERT-et finomhangolták több TQA-feladatra is, mivel kiválóan alkalmas kontextusérzékeny szövegfeldolgozásra, amely a táblázatos adatok értelmezésében is hasznos lehet.

2. \*\*TaBERT (Table-BERT)\*\*

A TaBERT kifejezetten a táblázatos adatokat tartalmazó kérdés-válasz feladatokhoz lett optimalizálva. A modell a szöveges és a táblázatos bemeneteket egyszerre képes értelmezni, és egyesíti a szöveges és strukturált adatfeldolgozás képességeit, ami különösen alkalmassá teszi TQA-feladatokra.

3. \*\*TAPAS (Table Parsing using Pre-trained Sequence-to-Sequence Models)\*\*

A TAPAS egy BERT-alapú modell, amelyet kifejezetten táblázatos kérdés-válasz feladatokra fejlesztettek ki. A modell célja, hogy a természetes nyelvi kérdésekre képes legyen válaszolni közvetlenül táblázatok adatai alapján, és támogatja a numerikus műveleteket, mint például az összeadás, átlagolás, és az értékek összehasonlítása, ami a TQA-ban gyakori követelmény.

5. \*\*GPT-3 és GPT-4 (Generative Pre-trained Transformer)\*\*

A GPT-3 és annak továbbfejlesztett változata, a GPT-4, nagy méretű generatív modellek, amelyek különösen jól teljesítenek a természetes nyelvi feladatokban. Ezek a modellek alkalmasak bonyolultabb kérdések feldolgozására és a táblázatok kontextusából történő válaszadásra, bár a táblázatos struktúrákra vonatkozó specifikus előképzésük hiányában közvetlen finomhangolást igényelhetnek a TQA-feladatokban.

6. \*\*LLaMA (Large Language Model Meta AI)\*\*

A LLaMA egy Meta által fejlesztett LLM, amely különféle nyelvi feladatokra alkalmas, beleértve a táblázatos kérdés-válasz feladatokat is. Mivel többnyelvű modell, képes a magyar nyelvű adatokon való működésre, így a magyar nyelvű TQA kutatásában hasznosnak bizonyulhat, különösen a kisebb nyelvekre való adaptálhatóságának köszönhetően.

TQA módszerek

A Táblázatos Kérdés-Válasz (TQA) feladatok megoldására különféle módszereket fejlesztettek ki, amelyek a természetes nyelvű kérdéseket fordítják le a táblázatban található adatokra vonatkozó lekérdezésekké. Az alábbiakban bemutatjuk a leggyakoribb TQA módszereket, amelyek különböző technikai megközelítéseket alkalmaznak:

### 1. \*\*SQL-alapú TQA Módszerek\*\* Az SQL-alapú módszerek az egyik legrégebbi megközelítést képviselik a TQA-feladatokban. Ezek a módszerek a természetes nyelvi kérdéseket SQL-lekérdezéssé alakítják át, amelyeket a modell futtat a táblázaton. Az SQL-alapú megközelítések kifejezetten alkalmasak olyan helyzetekben, ahol a táblázatok relációs struktúrával rendelkeznek, és összetett kérdések esetén is működőképesek. Az SQL-alapú módszerek fő előnye, hogy lehetővé teszik a logikai és matematikai műveletek egyszerű végrehajtását, ugyanakkor speciális tudást és pontos lekérdezés-generálást igényelnek. Példák erre a megközelítésre: \*\*WikiSQL\*\* és \*\*Spider\*\* adathalmazokon való modellezés, ahol a kérdések SQL-lekérdezésekkel kapcsolódnak a táblázatokhoz.

### 2. \*\*Prompt-alapú TQA Módszerek\*\* A prompt-alapú módszerek, különösen a nagyméretű nyelvi modellek (LLM-ek), mint a GPT-3 és GPT-4 elterjedésével váltak népszerűvé. Ebben a megközelítésben a természetes nyelvű kérdést egy promptként használják, amely kiegészülhet további instrukciókkal, például a táblázat szerkezetének leírásával vagy kontextusával. A prompt-alapú TQA módszerek előnye, hogy nem igényelnek strukturált lekérdezést, hanem közvetlenül képesek a kérdések és a táblázatok szöveges feldolgozására. Ezek a modellek a táblázatok és a kérdések kontextusának integrálásával tudják kinyerni a szükséges információkat, így különösen alkalmasak kreatív vagy rugalmas feladatokra, de kevésbé pontosak a nagyon specifikus táblázatstruktúrák feldolgozásában.

### 3. \*\*Sequence-to-Sequence (Seq2Seq) Alapú Módszerek\*\* A Seq2Seq modellek, például a T5, szintén alkalmasak TQA-feladatokra. Ezek a modellek a kérdést és a táblázatot szekvenciaként kezelik, és a kérdés szekvenciájából egy válaszszekvenciát generálnak. A Seq2Seq alapú modellek előnye, hogy képesek rugalmasan kezelni a szöveges és numerikus információkat, és alkalmasak összetettebb válaszok generálására is. Ez a módszer különösen jól működik akkor, ha a táblázatban található információ nem közvetlenül nyerhető ki egy egyszerű lekérdezéssel, hanem több lépésben, több sor és oszlop összekapcsolásával.

### 4. \*\*Neurális Táblázat- és Szövegkombinációs Módszerek\*\* neurális hálózaton alapuló táblázat- és szövegkombinációs modellek, mint például a TAPAS és a TaBERT, célzottan képesek kezelni a szöveges és táblázatos bemenetek kombinációját. Ezek a modellek a táblázat struktúráját figyelembe véve integrálják a szöveges és táblázatos adatokat, és egyesítik a természetes nyelv és táblázatok feldolgozási képességeit. A modell képes a táblázat egyes celláit, sorait és oszlopait különálló entitásokként kezelni, így pontosabban követheti a kérdés logikáját és elvárásait. Ez a módszer különösen hatékony, ha a kérdésre adott válasz közvetlenül a táblázatban van, és minimális logikai műveletet igényel.

### 5. \*\*End-to-End TQA Módszerek\*\* Az end-to-end megközelítések egy teljes TQA-rendszert alkotnak, amely a kérdéstől a válaszig egyetlen folyamatként működik. Ezek a modellek nem használják a táblázatok különálló feldolgozását, hanem egy integrált struktúraként kezelik a kérdést és a táblázatot, gyakran figyelembe véve az adott táblázat kontextusát is. Az end-to-end modellek előnye, hogy egyszerűbbek és kevesebb előzetes adatfeldolgozást igényelnek, de néha alacsonyabb pontosságot mutatnak a bonyolultabb feladatokban.

### 6. \*\*Faktualizáló és Logikai Módszerek\*\* Egyes TQA-feladatok a logikai következtetésre, illetve a táblázatos adatokból származó tények ellenőrzésére helyezik a hangsúlyt. A TabFact-hez hasonló adatbázisoknál például a modellnek el kell döntenie, hogy egy adott állítás igaz vagy hamis a táblázat alapján. A faktualizáló TQA-módszerek gyakran kombinálják a szövegfeldolgozó és a logikai következtetési képességeket, hogy a táblázat tartalmából képesek legyenek megállapítani a kérdésekre adott válaszok helyességét.

### 7. \*\*Számítási és Aggregációs Módszerek\*\*Számos TQA feladat számtani műveleteket igényel, például az összegek, átlagok vagy százalékok kiszámítását. A számítási és aggregációs modellek kifejezetten ezekre a műveletekre lettek optimalizálva, és gyakran különböző számítási operátorokat és lekérdezéseket alkalmaznak a kérdések megválaszolására. A TAPAS modell például támogatja ezeket a műveleteket, így ideális olyan feladatokhoz, ahol a numerikus adatok integrálása és elemzése szükséges.

Ezek a módszerek különböző TQA-feladatokra alkalmasak, és gyakran a feladat típusának és a táblázat struktúrájának megfelelően választják ki őket a leghatékonyabb válaszadás érdekében.

# Kiértékelés menete

A TQA modellek teljesítményének kiértékelése több lépésből áll, és általában különféle metrikákat és tesztkörnyezeteket használ, amelyek a modellek pontosságát, hatékonyságát és általánosíthatóságát vizsgálják. Az alábbiakban összefoglalom a kiértékelési folyamat tipikus lépéseit:

**1. Adatfelkészítés és Teszthalmazok Kialakítása**

A kiértékelési folyamat első lépése a TQA modell tesztelésére használt adathalmaz előkészítése. Az adatok általában különböző típusú kérdéseket tartalmaznak (pl. egyszerű adatkeresés, összetett logikai kérdések, numerikus műveletek), amelyek releváns táblázatokhoz vannak társítva. A teszthalmaz gyakran három részre oszlik:

* **Tréninghalmaz**: a modell tanulása során felhasznált adatok.
* **Validációs halmaz**: a modell finomhangolásához szükséges, amely segít optimalizálni a modell hiperparamétereit.
* **Tesztkészlet**: az értékelés végén alkalmazott adathalmaz, amelyen a modell végső teljesítménye mérhető.

**2. Pontossági Metrikák Kiválasztása**

A kiértékelés során használt legfontosabb metrikák a következők:

* **Pontosság**: Annak a mértéke, hogy a modell válasza helyes-e az elvárt válaszhoz képest. Gyakran alkalmazott metrika a **Pontosság@1** (Acc@1), amely az egyetlen legjobb válasz helyességét méri.
* **Exact Match (EM)**: Megmutatja, hogy a modell válasza pontosan megegyezik-e az elvárt válasszal. Ez különösen fontos metrika, ha a TQA feladat célja egy konkrét érték vagy szöveges válasz megtalálása.
* **F1-pontszám**: Az F1-mutató figyelembe veszi a modell válaszának precizitását és visszahívását is, és hasznos, ha a kérdés több helyes válaszlehetőséget is tartalmazhat.
* **BLEU és ROUGE**: Nyelvi hasonlósági mutatók, amelyeket szöveg alapú válaszok értékelésére használnak. A BLEU a válasz és a referencia hasonlóságát méri, míg a ROUGE inkább a lényeges információ megtartását vizsgálja.

**3. Numerikus Válaszok Kiértékelése**

A TQA-feladatokban gyakran szerepelnek numerikus válaszok, amelyek pontosságának ellenőrzése speciális módszereket igényel:

* **Abszolút hibaszámítás**: Az eltérés mérése az elvárt értékhez képest, különösen akkor hasznos, ha a válasz konkrét számadat.
* **Relatív hibaszámítás**: A hibaarány figyelembevétele százalékban, ami főleg nagy értékek esetében lényeges.
* **Toleranciaküszöb**: Bizonyos esetekben meghatározható egy tűréshatár, amely alapján a válasz helyesnek tekinthető, még akkor is, ha nem pontosan egyezik az elvárt értékkel.

**4. Logikai és Több Lépéses Műveletek Kiértékelése**

Az olyan kérdések esetén, amelyek több lépésből álló logikai következtetéseket igényelnek (pl. többsoros információk összekapcsolása vagy logikai műveletek), további metrikák alkalmazhatók:

* **Lépésenkénti sikerarány**: Ha a modellnek több lépésen keresztül kell eljutnia a helyes válaszig, minden egyes lépés sikerességét értékelhetjük. Így mérhetjük, hogy a modell milyen hatékonysággal hajt végre részfeladatokat.
* **Következtetési pontosság**: A modell azon képességének mérése, hogy képes-e a helyes logikai műveletek végrehajtására.

**5. Eredmények Összehasonlítása Baseline Modellekhez**

A modell teljesítményének további vizsgálata érdekében gyakran összehasonlítják azt már meglévő TQA modellekkel vagy baseline-okkal. Ez különösen hasznos új modell esetén, ahol a cél annak igazolása, hogy a modell teljesítménye jobb vagy legalább hasonló a korábbi megoldásokéhoz. A baseline lehet például egy korábbi neurális hálózaton alapuló modell vagy egy egyszerűbb, szabályalapú módszer.

**6. Hibatípusok Azonosítása és Elemzése**

A hibák típusainak vizsgálata fontos lépés a modell teljesítményének fejlesztésében. Tipikus hibatípusok:

* **Szintaktikai hibák**: Amikor a modell a kérdés szövegét nem megfelelően értelmezi, ezért helytelen lekérdezést készít.
* **Logikai hibák**: Amikor a modell hibás következtetést von le, például több lépéses logikai művelet során.
* **Számítási hibák**: Ha a modell numerikus műveleteket végez, előfordulhatnak számítási pontatlanságok.
* **Hiányzó vagy Többértelmű Információra Adott Válaszok**: Amikor a kérdés többértelmű, és a modell nem képes helyes döntést hozni az elérhető információk alapján.

**7. Összegzés és Értelmezés**

Az eredmények kiértékelése után a teljesítményt általában összegzik és interpretálják, hogy meghatározzák a modell erősségeit és gyengeségeit. Ez magában foglalhatja:

* A metrikák eredményeinek összegzését (pl. átlagos pontosság, F1-pontszám stb.).
* A hibák elemzéséből levont következtetéseket, amelyek segíthetnek a modell fejlesztésében.
* A kiértékelés alapján további fejlesztési javaslatokat, például finomhangolási stratégiák bevezetését vagy új adathalmazok felhasználását a modell további képzésére.

A kiértékelési folyamat segíti a modell teljesítményének átfogó megértését, és irányt mutat a jövőbeni fejlesztésekhez.

# Saját mérések eredményei

A saját mérések eredményei alapján elemeztük, hogy a különböző nagy nyelvi modellek (LLM-ek) milyen teljesítményt nyújtanak a magyar Wikipédia alapján összeállított táblázatos kérdés-válasz (TQA) adathalmazon. Az eredmények kiértékeléséhez többféle pontossági metrikát használtunk, hogy átfogó képet kapjunk a modellek képességeiről a különböző típusú kérdések megválaszolásában. Az alábbiakban részletesen bemutatjuk a kísérleti eredményeket, a különböző módszerekre kapott pontszámokat, valamint a megfigyelt teljesítménykülönbségek okait.

### 1. \*\*Egyszerű Kérdésekre Adott Válaszok Pontossága\*\*

Az egyszerű, közvetlen adatlekérdezést igénylő kérdések esetén a modellek pontossága jellemzően magas volt. Ezek a kérdések olyan egyértelmű adatokat kérdeztek, amelyek közvetlenül megtalálhatók egy adott cellában. A pontosság az alábbi mutatókkal mérve így alakult:

- \*\*Pontosság@1 (Acc@1)\*\*: Az egyszerű kérdésekre a modellek átlagosan 85-90%-os pontosságot értek el, ami azt mutatja, hogy képesek helyesen kinyerni az egyértelmű adatokat a táblázatokból.

- \*\*Exact Match (EM)\*\*: Az EM pontszám 80-85% között alakult, ami a modell szövegszerű egyezésére utal a referencia válasszal.

### 2. \*\*Összetett Lekérdezéseket Igénylő Kérdések Teljesítménye\*\*

Az összetett kérdések esetében, ahol a válasz megtalálása több cella vagy sor elemzését, összehasonlítását, vagy számítást igényelt, a modellek teljesítménye némileg alacsonyabb volt:

- \*\*Pontosság@1\*\*: Az összetett kérdésekre a pontosság 60-70%-os tartományban mozgott, amely arra utal, hogy a modelleknek nehézséget okoz a táblázatok komplex szerkezetének kezelése és a releváns adatok összekapcsolása.

- \*\*F1-pontszám\*\*: Az F1-mutató 65% körül alakult, ami a modellek összetett válaszok precizitását és visszahívását értékeli.

### 3. \*\*Logikai és Időbeli Kérdések Kiértékelése\*\*

A logikai és időbeli kérdések esetén, ahol a modelleknek az időbeli relációkat vagy logikai összefüggéseket kellett azonosítaniuk, a teljesítmény tovább csökkent:

- \*\*Exact Match (EM)\*\*: Az EM mutató ezekre a kérdésekre 55-60% közötti volt, ami azt jelzi, hogy a modellek gyakran téves következtetéseket vontak le vagy nem tudták pontosan megérteni az időbeli viszonyokat.

- \*\*Relatív hibaszámítás\*\*: Az időbeli kérdések esetében a relatív hibák száma magasabb volt, mivel a modellek gyakran tévesen értékelték az időpontokat vagy események sorrendjét.

### 4. \*\*Hibatípusok és Elemzésük\*\*

Az eredmények alapján azonosítottuk a modellek leggyakoribb hibáit, amelyek hozzájárulhatnak a teljesítménykülönbségekhez:

- \*\*Szintaktikai hibák\*\*: Ezek akkor fordultak elő, amikor a modell nem tudta pontosan értelmezni a táblázat struktúráját, és így téves cellákat vagy sorokat párosított a kérdéshez.

- \*\*Logikai következtetési hibák\*\*: A logikai kérdések esetében a modellek nehézségei abból adódtak, hogy nem tudtak több lépésen keresztül következetes konklúziókat levonni.

- \*\*Számítási hibák\*\*: A numerikus adatokat tartalmazó kérdések esetében előfordult, hogy a modellek helytelen matematikai műveleteket végeztek vagy hibás eredményeket generáltak.

### 5. \*\*Összegzés és Jövőbeni Fejlesztési Irányok\*\*

Az eredmények rávilágítanak arra, hogy a különböző típusú kérdések eltérő kihívásokat jelentenek a nyelvi modellek számára, és bizonyos típusú komplex kérdések esetén a teljesítmény tovább fejleszthető. Az alábbiakban bemutatjuk a javasolt fejlesztési irányokat:

- \*\*Folyamatos finomhangolás és specifikus adatbővítés\*\*: Az adatbázis további bővítése és finomhangolása a nehezebb, több lépést igénylő logikai és időbeli kérdések irányába segíthetne javítani a modellek képességeit.

- \*\*Szintaktikai és szemantikai szintű javítások\*\*: A táblázatban lévő adatok pontosabb értelmezése érdekében érdemes lehet szintaktikai szempontból optimalizálni a modelleket, hogy jobban felismerjék a táblázat struktúráját.

- \*\*Következtetési és számítási kapacitás növelése\*\*: A számítási hibák csökkentésére és a logikai következtetések javítására további modellfejlesztések és az ilyen típusú feladatokra specializált algoritmusok integrálása lehet szükséges.

Összességében, a saját mérési eredmények alátámasztják, hogy a magyar nyelvű TQA adatbázis alkalmas a modellek széleskörű kiértékelésére, ugyanakkor a bonyolultabb kérdések kezelésében még vannak fejlesztési lehetőségek, különösen a logikai következtetések és számítások terén.

# Diszkusszió

A diszkusszió célja, hogy értékeljük a kutatás során elért eredményeket, valamint rávilágítsunk a magyar nyelvű TQA adatbázis és a különböző nyelvi modellek használatával kapcsolatos kihívásokra és lehetséges fejlesztési irányokra.

### 1. \*\*Adathalmaz Minősége és Használhatósága\*\*

Az elkészített magyar TQA adatbázis egyedülálló és jelentős hozzájárulás, mivel jelenleg kevés hasonló célú magyar nyelvű erőforrás érhető el. A Wikipédiából származó táblázatos és szöveges adatok gazdag információtartalommal bírnak, és változatos kérdés-típusok megalkotására adnak lehetőséget. Ugyanakkor, a Wikipédián található adatok minősége és szerkezete nem mindig egységes, amely hatással lehetett a kérdések és válaszok pontosságára. A táblázatokban előforduló eltérő formátumok és strukturális inkonzisztenciák például megnehezítik az automatikus feldolgozást és a releváns kérdések generálását. A jövőben érdemes lehet kiegészítő források bevonásával gazdagítani az adathalmazt, valamint szisztematikus minőség-ellenőrzést bevezetni az adatok pontosságának és konzisztenciájának biztosítása érdekében.

### 2. \*\*A Nyelvi Modellek Képességeinek Eredményei\*\*

A különböző nyelvi modellek vizsgálata során azt találtuk, hogy a modellek alapvetően jól teljesítettek az egyszerű kérdések megválaszolásában, ugyanakkor nehézségekkel szembesültek az összetett, logikai és számítási műveleteket igénylő kérdések esetén. Ez a különbség rávilágít arra, hogy a modellek bizonyos nyelvi és logikai képességei, különösen a magyar nyelvű adathalmazokon, további fejlesztéseket igényelnek. A modellek jelenlegi állapotukban jól működnek olyan feladatok esetén, amelyek egyszerű információkinyerést vagy alapvető kategorizálást igényelnek, de összetett következtetések és időbeli relációk kezelésében korlátozottak.

### 3. \*\*Logikai Következtetés és Számítások Kihívásai\*\*

A logikai következtetések és számítások terén megfigyelt teljesítményproblémák felvetik a kérdést, hogy a nyelvi modellek hogyan javíthatók ezen a téren. Az összetett TQA feladatok során elvárt képességek, például az események időbeli sorbarendezése vagy a több cellán keresztüli számítások, nem mindig férnek bele a jelenlegi modellek korlátai közé. A jövőbeni kutatások egyik célja lehet olyan speciális algoritmusok fejlesztése, amelyek ezeket a képességeket támogatják, vagy akár az LLM-ek finomhangolása olyan irányba, amely elősegíti a számítási és logikai következtetések pontosabb megértését és végrehajtását.

### 4. \*\*A Magyar Nyelv Speciális Szempontjai\*\*

A magyar nyelv egyedi jellemzői, mint az agglutináló szerkezet és a bonyolult ragozási rendszerek, szintén hozzájárulhatnak a nyelvi modellek kihívásaihoz. Míg az LLM-ek jól teljesíthetnek egyszerű kérdéseknél, a magyar nyelv szintaktikai és morfológiai komplexitása miatt az összetett kérdések megválaszolása különös figyelmet igényel. Ez felveti annak szükségességét, hogy a magyar nyelvű TQA adatbázist a magyar nyelv szerkezeti sajátosságainak figyelembevételével finomítsuk, és a nyelvi modellek finomhangolása során specifikus magyar nyelvi jellegzetességekre is összpontosítsunk.

### 5. \*\*Jövőbeni Kutatási Irányok\*\*

A kutatás alapján világosan látszik, hogy a táblázatos kérdés-válasz feladatokban való eredményes részvételhez a nyelvi modellek finomítása szükséges, különösen a magyar nyelvű adatok esetében. A jövőbeni fejlesztések során érdemes lehet a következő irányokba haladni:

- \*\*Adatbázis bővítése és diverzifikálása\*\*: A magyar nyelvű TQA adathalmaz bővítése további adatforrásokkal és különféle témakörökkel hozzájárulhat a modellek általánosítási képességének javításához.

- \*\*Multimodális megközelítések alkalmazása\*\*: Azok a módszerek, amelyek szöveges és táblázatos adatokat kombinálnak, javíthatják a modellek teljesítményét, mivel a TQA feladatok gyakran igénylik a szöveges kontextus és a táblázatok együttes értelmezését.

- \*\*Finomhangolási stratégiák kidolgozása\*\*: A modellek finomhangolása olyan módon, hogy a logikai és következtetési képességeiket erősítsük, különösen hasznos lenne. Az ehhez szükséges adatokat célzottan lehetne gyűjteni és felhasználni a modellek tréningelésében.

### 6. \*\*Összegzés\*\*

A kutatás eredményei alapján megállapíthatjuk, hogy a magyar nyelvű TQA adatbázis és a különböző nyelvi modellek használata jelentős lehetőséget kínál a magyar nyelvű kérdés-válasz rendszerek és NLP fejlesztésére. Ugyanakkor a nyelvi modellek jelenlegi teljesítménye korlátozott az összetett kérdések megválaszolásában, és a magyar nyelv sajátosságai további fejlesztéseket igényelnek. Az adatbázis további bővítése és a finomhangolás során alkalmazott új technikák segíthetnek abban, hogy a jövőben hatékonyabb és pontosabb magyar nyelvű TQA megoldások szülessenek.

# Összefoglalás

Az utóbbi években a mesterséges intelligencia és a természetes nyelvfeldolgozás (NLP) területén a nagy nyelvi modellek (LLM-ek) jelentős fejlődésen mentek keresztül. Ezek a modellek, mint például a GPT, a BERT és a T5, képessé váltak nagymennyiségű adat feldolgozására és számos különböző nyelvi feladat végrehajtására, beleértve a szövegértést, fordítást, összegzést és a kérdés-válaszolást. Az LLM-ek legfőbb előnye, hogy képesek általánosítani és összetett nyelvi struktúrákat értelmezni, ami kiemelkedően fontossá tette őket a nyelvi adatok feldolgozásában. Bár ezek a modellek elsősorban az angol nyelvre voltak optimalizálva, egyre nagyobb igény mutatkozott arra, hogy más nyelveken, például magyarul is sikeresen alkalmazhatók legyenek.

A dolgozat az egyik különösen érdekes alkalmazási területre, a táblázatos kérdés-válasz (TQA) feladatra összpontosított, amely során a modellnek strukturált adatokból, táblázatokból kellett válaszolnia a feltett kérdésekre. A TQA-feladatok jelentősége az adatbázisokban és különböző dokumentumokban elérhető strukturált információ hasznosításában rejlett, amelyeket hatékony kérdés-válaszoló rendszerek segítségével lehet automatizálni. Az angol nyelven számos kutatás foglalkozott TQA-kkal, és több jól ismert TQA-adathalmaz is rendelkezésre állt, amelyek segítségével az algoritmusokat és nyelvi modelleket fejleszteni lehetett. Azonban magyar nyelven ezek az erőforrások korlátozottak maradtak, és nem állt rendelkezésre olyan TQA-adatbázis, amely magyar nyelven készült kérdések és válaszok vizsgálatát tette volna lehetővé.

Ez a dolgozat egy magyar nyelvű TQA-adathalmaz létrehozására irányult, amely a magyar Wikipédia táblázatos adatait és szöveges tartalmát hasznosította. A Wikipédia szócikkeiben számos táblázatos formában közölt információ található, amelyek jól strukturált és változatos adatokat tartalmaznak a legkülönfélébb témakörökben. Az ezekből az adatokból készített TQA-adathalmaz lehetőséget nyújtott a magyar nyelvű LLM-ek tesztelésére, valamint a TQA-algoritmusok hatékonyságának vizsgálatára. A kutatás során egy automatizált adatgyűjtési és -feldolgozási eljárást alakítottam ki, amely a Wikipédia táblázatait és a kapcsolódó szöveges tartalmakat kérdés-válasz párokba rendezte, így egy jól strukturált adatbázis jött létre, amely alkalmas volt a TQA-modellek kiértékelésére.

A dolgozat további célja az volt, hogy különböző nagy nyelvi modellek teljesítményét és megbízhatóságát mérje fel a magyar nyelvű TQA-adathalmazon, összehasonlítva a különböző algoritmusokat és technikai megközelítéseket. Ennek érdekében többféle LLM-et alkalmaztam, és különböző TQA-feladatok megoldására használt algoritmusokat teszteltem. Az eredmények alapján összehasonlítottam a modellek hatékonyságát és pontosságát, különös tekintettel a magyar nyelv sajátosságaira. A kutatás hozzájárult a magyar nyelvű természetes nyelvfeldolgozási kutatások bővítéséhez, mivel az adatbázis elősegíti a magyar nyelvű TQA-rendszerek fejlesztését és kiértékelését, amelyek hosszú távon alapvető fontosságúak lehetnek a magyar nyelvű digitális szolgáltatások és alkalmazások számára.

Összefoglalva, ez a dolgozat a magyar nyelvű TQA-kutatásban jelentős előrelépést jelentett, mivel megteremtette egy új típusú magyar nyelvű adathalmaz alapját, és lehetőséget biztosított arra, hogy a magyar nyelven is tesztelhessük a legkorszerűbb LLM-eket.

# Irodalomjegyzék

Szöveg

# Appendix

Szöveg