

Tűzvédelmi szakértői rendszer létrehozása nagy nyelvi modellek segítségével

Fire Safety Expert System Development with Large Language Models

DOI: [HTTPS:// DOI.ORG/10.53793/RV.2024.2.3](https://doi.org/10.53793/RV.2024.2.3)

Absztrakt

Ez a kutatás a nagy nyelvi modellek katasztrófavédelmi célú alkalmazását vizsgálja. A cél egy olyan felület létrehozása, ahol tűzvédelemmel kapcsolatos kérdéseket lehet feltenni egy mesterséges intelligenciának magyar nyelven. A kutatás során jogi dokumentumokat és szakkönyveket használtam fel adatforrásként. A Nyelvtudományi Kutatóintézet által fejlesztett magyar nyelvű Puli GPT-3SX alapmodellt QLoRA módszerrel finomhangoltam az adatforrásokat felhasználva. A kutatás során arra a következtetésre jutottam, hogy lehetséges alacsony erőforrás igényű, hatékony chatbotot létrehozni, amely összefüggő és releváns válaszokat ad tűzvédelmi kérdésekre.

KULCSSZAVAK: MESTERSÉGES INTELLIGENCIA, NAGY NYELVI MODELL, KATASZTRÓFAVÉDELEM, TŰZVÉDELEM, CSEVEGŐ ROBOT

Abstract

This research investigates the application of large language models for disaster management purposes. The goal is to create an interface (chatbot) where questions about fire safety can be asked to an artificial intelligence in Hungarian. The research used Hungarian legal documents and professional books as data sources and fine-tuned the Puli GPT-3SX model of the Research Institute for Linguistics with the QLoRA method. The research suggests that it is possible to create an effective chatbot with low resource requirements that provides coherent and relevant answers to fire safety questions.

KEYWORDS: ARTIFICIAL INTELLIGENCE, LARGE LANGUAGE MODEL, DISASTERMANAGEMENT, FIRE SAFETY, CHATBOT

Bevezetés

A mesterséges intelligencia, azon belül is a gépi tanulással kapcsolatos kutatások fókuszja napjainkban egyre inkább a természetes nyelvfeldolgozással (NLP¹) kapcsolatos elemzések irányába mozdul el.

A nagy nyelvi modellek (LLM²) igazi áttörése a természetes nyelvek megértésével (NLU³) kapcsolatos. Az NLU az NLP egyik részterülete, amely a szövegértésre és a szemantikai elemzésre összpontosít. A nyelv megértése azt jelenti, hogy képesek vagyunk kommunikálni a számítógéppel az anyanyelvünkön, azaz egy természetes nyelven, nem csak programozási nyelveken keresztül. Ez az előrelépés olyan forradalmi változásokat hozhat el életünkben, mint annak idején a könyvnyomtatás vagy éppen az internet elterjedése.

Az a tény, hogy a számítógép bizonyos szinten már megérti a természetes nyelveket teljesen új perspektívákat ad a jövőre nézve. Számos munkahely, beosztás át fog alakulni, némelyek talán el is tűnnek, de újabbak jelenhetnek meg.

Jelen dolgozatban egy a ChatGPT-hez hasonló, annál jóval kisebb, magyar modellen alapuló saját fejlesztésű modellt szeretnék bemutatni, amely kifejezetten tűzvédelmi, katasztrófavédelmi kérdésekben jártas.

A modell neve: KATasztrófavédelmi Aszisztens, azaz KATA.

¹ Natural Language Processing

² Large Language Model

³ Natural Language Understanding

Háttér

Mit is értünk nyelvi modell alatt? Amikor számítógépek segítségével próbálunk meg leírni egy folyamatot, akkor modelleket készítünk, amelyek reményeink szerint egy elvárt viselkedést mutatnak. A nyelvi modell egy olyan valószínűségi eloszlás a szavak sorozatai között, ahol a modell minden egyes szóhoz valószínűségi értéket rendel egy szekvenciában, azaz a szövegben a következő szót kell előre jeleznie az előtte meglévő szavak alapján. A nagy nyelvi modell már olyan típusú nyelvi modell, amelynek háttérében egy neurális hálózat van. Ezek a neurális hálózatok az információkat nagyszámú paramétereikben (számok) tárolják. A nagy nyelvi modellek esetén ezeknek a paramétereknek a száma több milliárdos nagyságrendet is elérhet. A modell képzéséhez pedig nagy mennyiségű szöveges adatra van szükség.

A következőkben az architektúra és modell fogalmakat szinte szinonimként fogom használni. Az architektúra alatt azt a technikai leírást értjük, amely alapján felépül a teljes rendszer, elsősorban neurális hálózati elemekből. A modell alatt pedig egy konkrét architektúrára épített és betanított rendszert értünk.

A kezdetek

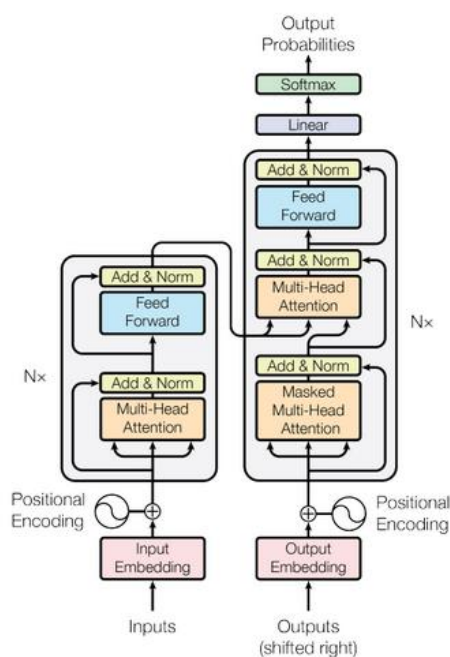
A nyelvi modellek története az 1960-as évekig nyúlik vissza. 1967-ben az MIT⁴ egyik professzora megépítette az Eliza nevezetű programot a természetes nyelv

megértésére. A program mintaillesztési és helyettesítési technikákat használt a nyelv megértéséhez (Weizenbaum 1966). A nyelvi modellezés a neurális hálózatok megjelenése előtt leginkább a szabályalapú, statisztikai módszerekre épült. A szabályalapú megközelítések mögött meghúzódó alapötlet az, hogy a nyelvben vannak olyan szabályok, amelyek arra vonatkoznak, hogy a szavak hogyan kapcsolódnak egymáshoz. Később a rekurrens neurális hálózatok (RNN⁵) megjelenésével a nyelvi modellezés gyors fejlődésnek indult. Az RNN-ek olyan neurális hálózatok, amelyek képesek figyelembe venni a betáplált információk (szavak) sorrendjét. Az RNN-ek rendelkeznek memóriával és képesek egy szekvenciát átalakítani egy másik szekvenciára.

A transzformer modell

2017-ben egy igazi áttörés történt a nyelvi modellek kutatása terén, megjelent az „Attention Is All You Need” (Vaswani 2017), „A figyelem minden, amire szükség van” című kutatás, és elhozta a transzformer architektúrát, amely egy teljesen új megközelítést alkalmazott.

Az 1. sz. ábra bemutatja a transzformer architektúra felépítését. Az ábra baloldalán található az úgynevezett kódoló egység, a jobb oldalon pedig a dekódoló egység. A transzformer architektúra szétválasztja a kódolást és a dekódolást, valamint többretegű figyelemalapú hálózatokat alkalmaz a nyelvi fordításra.



1. sz. ábra: Transzformer architektúra

Forrás: <https://arxiv.org/pdf/1706.03762.pdf> [Letöltés: 2023.07.10.]

⁴ Massachusetts Institute of Technology

⁵ Recurrent Neural Network

A transzformerek a bemeneti tokenek⁶ sorozatának feldolgozását végzik úgy, hogy minden egyes tokenhez olyan reprezentációt számítanak ki, amely a teljes bemeneti szekvencia kontextusában rögzíti annak jelentését.

Kódoló típusú modellek

A kódoló típusú nagy nyelvi modellek olyan gépi tanulási modellek, amelyek szövegkódolási feladatokra összpontosítanak, azaz szöveg-bemenetet alakítanak át egy kódolt reprezentációvá, amely a szöveg jelentését és tartalmát hordozza. A legismertebb és legelső ilyen architektúra a BERT⁷, egy olyan típusú nagy nyelvi modell, amelyet a Google kutatói fejlesztettek ki. A BERT forradalmi előrelépést jelentett a természetes nyelvfeldolgozás (NLP) terén.

A nyelvi modelleket az előzetes tanítás során hatalmas mennyiségű nyelvi adattal tréningezik. Ez a folyamat arra kényszeríti a modellt, hogy a szövegrészletekben rejlő összefüggéseket és kontextusokat megértse.

A BERT nagy előnye, hogy kontextuális reprezentációkat készít a szavak számára, azaz a szó jelentése és reprezentációja attól függ, hogy az milyen szöveggörnyezetben található. 2019-től ezt a modellt beépítették a Google keresőjébe is (Devlin et al. 2018).

Dekódoló típusú modellek

A legismertebb ilyen modell a GPT⁸ nevezetű nagy nyelvi modell, amelyet az OpenAI fejlesztett ki (Radford 2018).

A GPT modell előzetes tanítása szintén hatalmas mennyiségű szöveges adattal történik. A képzés során a modellnek mindig a következő szót kell kitalálnia, tehát a képzés alatt a modellnek nincs információja a kitalálendő szó utáni szavakról. A képzés során a modell megtanulja a nyelv mintázatait, és reprezentációkat fejleszt ki a szavak és a szöveggörnyezet közötti kapcsolatokra. A GPT előnye, hogy kontextuális reprezentációkat készít az előző szövegrészletek figyelembevételével, így képes értelmes és koherens szöveg generálására.

Kódoló-dekódoló típusú modellek

A kódoló-dekódoló típusú modellek a transzformerekről szóló cikkben (Vaswani 2017) megjelent eredeti architektúrát valósítják meg. Ebben az architektúrában a bemenetet először a kódoló átalakítja egy kódolt vektor reprezentációvá, amit átad a dekódoló modulnak, amely azt visszaalakítja kimenetűvé. Leggyakoribb felhasználási eset a nyelvek közötti fordítók megvalósítása, ahol a bemeneti nyelven betáplált szekvenciáról készül egy kódolt reprezentáció, amit majd a dekódoló vissza alakít a célnyelv formátumára.

Transzformer modellek képzése

A nagy nyelvi modellek képzése rendkívül költségigényes feladat, ezért nagyon kevés cég képes arra, hogy saját modellt fejlesszen. A fejlesztési költségek a modell növekedésével (paraméter számának növekedésével) skálázódnak. A BLOOM⁹ nevezetű 176 milliárd paraméterrel rendelkező nyelvi modell képzése nagyjából 118 napig tartott és 433196 kWh áramfogyasztással járt. A képzést 384 NVIDIA A100 GPU-val 48 számítási egység segítségével végezték el (Luccioni et al. 2022). Magyarországon jelenleg a Komondor nevű, pilot státuszban lévő szuperszámítógép képes csak hasonló feladatok elvégzésére. A Komondor specifikációja szerint a GPU partíciója 232 db NVIDIA A100-as GPU-val rendelkezik, és a fenti BLOOM modell képzése számításaim szerint több, mint 190 napig tartana, amennyiben a szuperszámítógép csak a nagy nyelvi modell képzésével foglalkozna.

Hazánkban jelenleg a legnagyobb nyilvánosan elérhető és magyar nyelvű adatokon képzett modell a Nyelvtudományi Kutatóközpont által készített, 6,7 milliárd paraméteres PULI-GPT-3SX (Yang Z. G. 2023).

Ezek a betanított modellek még alapmodellnek tekinthetőek, tehát a következő szavakat tudja megjósolni egy szövegbemenet, mint kontextus folytatásaként. Az ilyen modelleket prompt programozással már lehet használni, azaz néhány bemeneti példával képes megoldani nyelvtechnológiai feladatokat. A modell feladat specifikus képzéséhez azonban finomhangolásra van szükség, amely esetén a modell képes lehet kifejezetten tartományi ismeretekre szert tenni, tehát pl. orvosi kérdésekre képes lehet válaszolni, vagy akár tűzvédelmi kérdésekre is.

6 szavak, szótöredékek, karakterek

7 Bidirectional Encoder Representations from Transformers, azaz kétirányú kódoló ábrázolás transzformerből

8 Generative Pre-trained Transformer, azaz generatív előképzett transzformer

9 BigScience Large Open-science Open-access Multilingual Language Model (BLOOM)

Megközelítés

Az elérendő cél, hogy olyan modellt képezzünk ki, amely egy adott témában, esetünkben a tűz- és katasztrófavédelmi kérdésekben képes legyen releváns és koherens választ adni. Amikor emberekről beszélünk, akkor az adott témában jártas embereket szakértőnek nevezzük. Ebből kifolyólag a célunk egy szakértői rendszer készítése, amely humán közreműködés nélkül képes adekvát válaszokat adni.

A probléma megközelítése során két fő döntést kellett meghozni. Az első az architektúra és a modell kiválasztása, a második a képzési adatok meghatározása, gyűjtése. Ezen döntések stratégiai fontossággal bírnak, hiszen a végeredmény, az elkészült modell teljesítménye nagyrészt ezeken múlik.

A modellválaszték bemutatása

A modellek képzési részében leírtak alapján egyértelmű, hogy teljesen új alapmodellt nem tudunk készíteni jelentős anyagi ráfordítások nélkül, bár ez lenne a legoptimálisabb, hiszen ilyenkor már az alapmodell tanításához használt adatok közé be lehet illeszteni a tartományi tudást reprezentáló adatokat. Ilyenek lehetnek a témával kapcsolatos jogszabályok, tankönyvek, elemzések és szinte bármilyen jó minőségű szöveges anyag, amely relevanciával bír a tűz- és katasztrófavédelemmel kapcsolatosan.

Az interneten több alapmodell is elérhető különböző licenc szabályokhoz kötve, ezek nagy része CC BY-NC 4.0¹⁰ licenccel rendelkezik, azaz megosztható, módosítható, de kereskedelmi célokra nem használható fel. Mivel a projekt kutatási fázisban van, ezért ez a licenc megfelelő lehet. Létezik néhány modell, ami CC BY-SA 2.0¹¹ licenc alatt van, azaz szabadon felhasználható kereskedelmi célokra is, de ezek között nem találtam olyat, ami jelentős magyar nyelvű szövegen volt képezve.

Az első döntés a modell kiválasztásakor a modellméret kell, hogy legyen, hiszen a modellmérethez kell rendelni a szükséges hardver erőforrásokat, amelyek azonban csak korlátozottan álltak rendelkezésemre. Általánosságban elmondható, hogy a dekóder modellek mérete néhány típusra korlátozódik a paraméterszámot illetően.

Az elérhető modellek sora a legtöbb esetben a következő: 7B, 13B, 30B, 34B, 40B, 65B, 70B, 176B, ahol a szám a milliárd paramétert jelöli, a B az angol „billion” szóra utal, ami milliárdot jelent. Az egyik nagy közösségi oldalt is működtető Meta cég két modell családot is elkészített Llama, Llama-2 néven, az utóbbiból a 7B, 13B, 70B modellméreteket adták ki, a 34B-t későbbre

időzítették. A jelenleg használatos modellek legtöbbje ezekre az ingyenesen elérhető Llama, illetve Llama-2 modellekre

épít.

A nagyobb modellek rendre sokkal jobb eredményeket érnek el a teszteken (Touvron et al. 2023b), azonban nagyobb erőforrás szükséges működtetésükhöz, és lényegesen nagyobb a további képzéséhez, az úgynevezett finomhangoláshoz.

A modellek fontos paramétere az egyszerre feldolgozható kontextusméret, ez általában 2k, azaz kétezer token, de az újabb modellek esetében előfordulnak ennél nagyobb méretek is. A kontextusban szereplő szövegen (tokenek sorozata) tanul a modell egy-egy lépésben a tanulási folyamat során. A nagyobb kontextusméret azt jelenti, hogy a modell képes megragadni és megtalálni egymástól jóval távolabb eső szövegrészeket, majd összekapcsolni azokat.

A rendelkezésemre álló erőforrások csak a legkisebb, azaz a 7B modellek finomhangolását tették lehetővé, ezért ezeket vizsgáltam elsősorban.

A modell kiválasztása

Három modellt teszteltem le és hasonlítottam össze, amelyekből kiválasztottam a véglegeseket.

Llama 7B modell

A modellt a Meta adta ki, 1 T (trillió) 10^{12} tokenre képezték ki, legnagyobb részt angol nyelven, de 20 egyéb nyelvű szöveg, köztük magyar is volt a képzési adathalmazban. A modell kontextusmérete 2k. A magyar nyelvű szövegek aránya kevesebb, mint 1% volt.

A modellt egy gyors finomhangolás során 20 ezer kérdés-válasz páron tanítottam, majd értékeltem. Mivel nem volt képes koherens, nyelvtanilag helyes mondatokkal válaszolni a feltett kérdésekre, ezért elvettem a modell használatát (Touvron et al. 2023a).

Llama-2 7B modell

A modellt a Meta adta ki a Llama 7B továbbfejlesztett változataként, ezt már 2T tokenre képezték ki, a kontextusméret itt már 4k volt. A magyar nyelvű anyagok aránya a teljes képzési korpuszon belül 0,03% volt (Touvron et al. 2023b). A modellt szintén finomhangoltam 20 ezer kérdés-válasz páron, majd értékeltem az eredményt. Az eredmény jobb volt a Llama 7B modellhez képest, ezért kiválasztottam, hogy a későbbiekben finomhangolva értékeljem az eredményeit.

¹⁰ Attribution-NonCommercial 4.0 International/

¹¹ Attribution-ShareAlike 2.0 Generic

PULI-GPT-3SX modell

A modellt a Nyelvtudományi Kutatóintézet készítette. A PULI-GPT-3SX-et 32,4 milliárd szavas korpuszon tanították be, a token számot nem adták meg, de kb. 2,5-3 szoros token mennyiséggel kell számolni, azaz 80-90 milliárd token, a kontextusméret 2k volt (Yang 2023). Ezt az alapmodellt szintén kiválasztottam, mert az előbbieknél jóval koherensebb, nyelvtanilag jobb szövegeket generált, köszönhetően a több magyar nyelvű képzési adatnak.

ChatGPT modell

Az OpenAI által létrehozott ChatGPT az egyik legfejlettebb nyelvi modell, jelenleg több változata is létezik, az első verzió a 3,5-ös verziónevet kapta és a GPT3-ra épülő finomhangolt verzió volt. Ezt 2022 decemberében adták ki. Jelenleg már elérhető a 4,0-ás modell is, amely elődjénél fejlettebb képességekkel bír, azonban az OpenAI nyilvánosan nem adott ki információkat a modell hátteréről.

Az OpenAI a ChatGPT-hez kapcsolódóan egyetlen modelljét sem adta ki nyilvánosan, azok csak webfelületen és API¹²-n keresztül érhetőek el. Ez annak a veszélyével jár, hogy a nyelvi modell használatakor minden, a modellnek átadott információ az OpenAI-hoz kerül. Ez a védelmi szférában teljességgel megengedhetetlen, de számos ipari szereplő részére is veszélyt jelenthet egy kódrészlet vagy egyéb bizalmas információ kikerülése az ellenőrzött körülmények közül.

Egy felmérés szerint (Coles 2023) a szellemi munkát végző dolgozók 10,8%-a próbálta ki már legalább egyszer a ChatGPT-t a munkahelyén, és 8,6%-a illesztett be vállalati adatokat is. A munkavállalók legtöbbször a saját munkájukat segítő, személyi asszisztensként használják a ChatGPT-t, amire alkalmas is, hiszen rendkívül fejlett képességekkel bír, azonban minden ilyen használat adatátadással jár. A felmérés szerint az alkalmazottak 4,7%-a illesztett be (adott át) bizalmas adatokat. A ChatGPT esetében a használati feltételek során deklarálják, hogy az átadott tartalmat felhasználhatják szolgáltatásaik javítására, erről külön formanyomtatványon kell leiratkozni.

Összegezve: a ChatGPT használata a védelmi szférában véleményem szerint egyáltalán nem javasolt, kivéve azokat az eseteket, ahol kizárólag teljesen nyilvános adatokat kell feldolgozni. Meg kell jegyeztem, hogy adatkészlet előállításához használtam a ChatGPT-t az API-ján keresztül, azonban a fentiek miatt teljesen elvettem a ChatGPT, mint modell használatát a szakértői rendszer létrehozásához.

A tanítóadatok kiválasztása

A tanítóadatok kiválasztásánál két fő szempontot tűztem ki. Az első, hogy a modell képes legyen megérteni bizonyos utasításokat és az utasításnak megfelelő válaszokat generálni. A második szempont a tartományi, azaz a tűzvédelmi, katasztrófavédelmi ismeretek elsajátítása volt. Mindkét szemponthoz ki kell választani az adatok körét, azokat át kell alakítani kérdés-válasz típusú formátumra.

Utasításokkal kapcsolatos adatkészlet

Több ilyen utasításkövető (instruction-following) adatkészlet létezik, amelyek közül a választásom a Stanford Alpaca adatkészletre esett, ennek is egy tisztított változatára. Ez egy szintetikus adatkészlet, tehát gépi úton létrehozott adatokról van szó, amit a text-davinci-003 nevű nagy nyelvi modell felhasználásával, 175 kiinduló utasítással, öninstrukció segítségével hoztak létre (Wang 2022).

Az adatkészlet 52 ezer kérdés-válasz párt tartalmaz, angol nyelven. A tisztítás azt jelenti, hogy eltávolították azokat az adatokat, amelyek helytelenek voltak. Mivel szintetikus adatokról beszélünk, ezért előfordultak olyan esetek, amikor az adatkészletet létrehozó modell hallucinált, ezért helytelen adatok is készültek.

Az adatkészletet a Google fordító API segítségével lefordítottam magyar nyelvre. A fordítás nem tökéletes, előfordulnak benne nem teljesen érthető vagy nyelvtani hibás adatok, de ezek manuális ellenőrzését idő hiányában nem is terveztem elvégezni. A hibás fordítási adatok valamelyest rontják a modell teljesítményét.

Az adatok utasítás, bemenet és kimenet kategóriák szerinti és JSON¹³ formátumba kerültek elmentésre.

Egy példa az adatkészletben található utasításokra:

utasítás: „Mi Franciaország fővárosa?”

bemenet: „”

kimenet: „Franciaország fővárosa Párizs.”

Fent a bemenet mező üres.

Egy ilyen kérdés-válasz pár megtanulása után azt várjuk a modelltől, hogy egy hasonló kérdésre is megfelelő választ tudjon adni, azaz „Mi Magyarország fővárosa?” kérdésre a kimeneten „Budapest”-et várunk el.

Egy másik példa, ahol a bemenet mező is szerepet kap:

utasítás: „Osztályozza a következőket állatokra, növényekre és ásványi anyagokra!”

bemenet: „Tölgyfa, rézérc, elefánt.”

kimenet: „Állatok: Elefánt Növények: Tölgyfa Ásványok: Rézérc.”

¹² Application Programming Interface

¹³ JavaScript Object Notation

Ebben a feladatban osztályozással kapcsolatos ismereteket szerez a modell, ti. az elefánt egy állat, a tölgyfa pedig növény.

A közel 52 ezer utasítás, bemenet és kimenet megtanulása után a modell általánosító képességei a tanult utasításokkal kapcsolatban javulnak.

Az utasításkövető adatkészlettel az volt a célom, hogy a modell felismerjen teljesen általános szándékokat és a szándéknak megfelelő kimenetet, eredményt generáljon.

A tanulási folyamat során a modell megkapja az utasítást, a bemenetet és az elvárt kimenetet, ezt megtanulja, majd amikor a modell elkészül, már csak az utasítást és a bemenetet kapja meg, a kimenetet pedig a modell állítja elő.

Tűzvédelmi adatkészlet

A tűz- és katasztrófavédelmi adatok összeállításánál három különböző adatforrást határoztam meg. Az első a jogszabályok kategória, a második kategóriát azok a szakanyagok, jegyzetek alkotják, amelyeket szintén az internetről töltöttem le, jellemzően a Nemzeti Közszolgálati Egyetem jegyzetei közül. A harmadik csoportot azok a tesztek alkotják, amelyeket kollégáim állítottak össze az elmúlt időszakban a tűzvédelemben dolgozó kollégák oktatására.

Az első két kategóriába tartozó szövegeket UTF-8¹⁴ formátumú szöveggé konvertáltam, majd az OpenAI API és egy Python script segítségével a szövegekből szintetikus kérdéseket és válaszokat generáltam. A dokumentumok feldolgozása végén több, mint 25 ezer tűz- és katasztrófavédelemel kapcsolatos kérdés-válasz párt generáltam.

A végső adatkészlet összeállítása

A végső adatkészlet létrehozásához az összes meglévő adatkészletet konkatenáltam, majd Huggingface Dataset formátumba konvertáltam, hogy a modell finomhangolása során a Huggingface által használt programozási könyvtárak segítségével a lehető legegyszerűbben tudjam az adatokat beolvasni és képzésre átadni a modellnek. A Huggingface egy programozási keretrendszer, amely főként a nagy nyelvi modellekre összpontosít.

Meg kell jegyezni, hogy a képzési adataim kevesebb, mint 1 %-a ember által készített kérdés-válasz pár, az adatok nagy része szintetikus előállított. Az adatok minősége nagyon fontos a képzés során, az ember által

gondosan összeállított adatkészletek segítségével sokkal jobb eredményeket lehet elérni akár kevesebb adattal is, a szintetikus adatokhoz képest (Zhou 2023). Ilyen nagyszámú adat azonban nem állt rendelkezésemre, illetve elkészítése rendkívül munka- és időigényes lett volna, ezért használtam főként szintetikus előállított adatokat.

A modellek finomhangolása

A modellek finomhangolása során az alapmodell részére további tudást, viselkedési formákat szeretnénk tanítani. A tudás alatt ténybeli információkat, azaz olyan szövegeket kell érteni, amelyeket az alapmodell még egyáltalán nem, vagy csak keveset látott, ezért ilyenkor lehetőség nyílik szak-specifikus adatok tanítására. A viselkedési formák tanítása alatt pedig azt értjük, hogy a modell képes legyen válaszolni kérdésekre, érvelésre és egyéb emberi viselkedésmintázatok utánozására. Ezt a képességet a már említett Alpaca adatkészlet biztosította.

A modellek finomhangolása már kevésbé költséges, mint az alapmodellek készítése, de ez is rendkívül hardverigényes lehet, ami a valóságban azt jelenti, hogy nagy GPU kapacitás szükséges és sok VRAM¹⁵.

A gyakorlatban több PEFT¹⁶ módszer létezik a költségek és az erőforrások csökkentéséhez, én a LORA¹⁷ módszert, azaz alacsony rangú mátrixok modellbe injektálását választottam (Hu et al. 2021). Ennek a megoldásnak az az előnye, hogy sokkal kevesebb számú paramétert kell tanítani a modell finomhangolása során. Az alapmodell súlyait lefagyasztjuk, azaz nem módosítjuk, csak a LORA adapterek súlyait kell megtanulni a modellnek, amely jóval kisebb feladat. Nagyobb modellek esetén még a LORA módszer is erőforrás-igényes lehet. Azért, hogy akár otthoni gépeken is elérhető legyen a módszer, kifejlesztették a QLORA¹⁸ eljárást, amely annyiban módosítja az eredeti LORA algoritmust, hogy az alapmodellt 4 bitesre kvantálják és az adapter tanítása 16 biten történik (Dettmers 2023).

A QLORA módszer segítségével tanítottam be mind a két modellt.

Mindkét modellt 3 korszakon keresztül tanítottam, azaz a teljes tanító adathalmazt háromszor láthatták. Ezt tapasztalati úton határoztam meg, úgy, hogy több modellt is készítettem, szám szerint 20 darabot, különböző paraméter beállításokkal, majd az általam legjobbnak ítéltet választottam a végső tesztelésre.

A modellek értékeléséhez kétféle módszert használtam. Az egyik során teszteket kellett kitöltenie a

¹⁴ 8-bit Unicode Transformation Format

¹⁵ Video Random-Access Memory

¹⁶ PEFT (Parameter Efficient Fine-Tune)

¹⁷ Low-Rank Adapters

¹⁸ Quantized Low-Rank Adapters

modellnek és azt értékeltem, míg a másik módszernél a szabadszöveges kérdésekre adott válaszokat vizsgáltam.

Tesztek kitöltése

A tesztek kitöltése során tűzvédelmi szakvizsga teszteket kellett megoldania a modelleknek. A feladatokat az interneten nyilvánosan is elérhető weboldalról töltöttem le és használtam fel az értékeléshez. A tesztek az 1-es foglalkozási ághoz, a hegesztők és az építőipari tevékenység során nyílt lánggal járó munkát végzőkhöz kapcsolódóak voltak. Ezek a feleletválasztós tesztek összesen 12 kérdést, kérdésenként 3 válaszlehetőséget – amelyek közül csak egy elfogadható – tartalmaztak.

A tesztekhez használtam egy rendszerpromptot, amely a modellnek ad egy alaphelyzet beállítást, hogy éppen milyen környezetben van. További promptként a kérdést és a lehetséges válaszokat adtam át. A modell a működése során szöveges információkat kap bemeneti adatként, ezeket hívjuk promptnak. A rendszerprompt, amit minden kérdéshez megkap a modell, választ segítő információkat tartalmaz. A modellnek a bemeneti adatok alapján kell kimenetet generálnia, ezt tanulja meg a képzések során. Minél több bemeneti adatot kap – ez képezi a kontextust a témához –, annál jobb válaszokat tud adni. A rendszerprompt tehát a modellnek egyfajta kontextusba helyezése, ami elmondja, hogy most tűzvédelem a téma és ilyen tesztet kell kitölteni. Ezt az információt egészítjük ki a teszt kérdéssel.

Nézzük meg egy példán keresztül ezt a folyamatot:

Rendszerprompt: „Ön egy tűzvédelmi szakértő és egy hegesztők és az építőipari tevékenység során nyílt lánggal járó munkát végzők részére kiállított tesztet kell kitöltenie.”

Itt a modell információkat kap arról, hogy tűzvédelem, hegesztők, nyílt láng stb. Ezeket az információkat a válasz generálásánál fel tudja használni.

Prompt: „### Instruction: Szabad-e az oxigénpalackot olajjal, zsírral átitatott textíliával tisztítani?

A. szabad, ha a gázpalack szelepe zárva van

B. szabad, így könnyebben működnek majd a szelepek

C. nem szabad!”

Az ### Instruction: kifejezéssel jelezzük a modell számára, hogy itt egy feladatot kap. A modell finomhangolásakor a feladatok, kérdések ilyen ### Instruction, ###Answer formában kerültek átadásra a modellnek, tehát megtanulta, hogy egy ilyen kifejezés után válaszolnia kell a kérdésre. De volt olyan modell is, amit ### Question, ###Answer mintával tanítottam. Az eredmény szempontjából nem releváns a formátum használata.

A modell válasza:

„###Answer: Nem, az oxigénpalackot nem szabad olajjal, zsírral átitatott textíliával tisztítani.”

Egy másik modell válasza az adott kérdésre:

„###Answer: C. Nem szabad! A gázpalackokkal kapcsolatban előfordulhat, hogy a gázpalackokat olajjal vagy zsírral átitatják, de ez nem minősül tisztításnak, hanem veszélyt jelenthet a biztonságra. Az ilyen típusú tisztítás nem végezhető el, mert a gázpalackon lévő anyagok felmelegedése robbanást okozhat. ...”

Mindkét válasz helyes, csak a szövegek hosszúságában van eltérés. Az, hogy a modell milyen hosszú választ adjon, egy állítható paraméter. Amennyiben hosszú választ kérünk gyakran előfordul, hogy a modell „elkalandozik”, vagyis kiesik a kontextusból és nem a témához szorosan kapcsolódó szöveget generál.

A szakvizsga teszten a legjobb eredmény 66%-os eredmény volt, 12 kérdésből 8 válasz volt megfelelő. Ez a való életben azt jelentené egy szakvizsgáztatás során, hogy szóbeli vizsgát kellene tenni a vizsgázónak.

A modellek tehát többnyire megértik a kérdést és ki tudják választani a helyes választ, ami rendkívül biztató egy viszonylag kis paraméterszámú modell esetén.

Kérdés-válasz párosok

A modelleket kérdésekkel és az arra adott válaszokkal is teszteltem. Vizsgáltam a modell „kreativitását”, tűzvédelmi szakmai ismereteit, helyesírását, koherenciáját.

A tapasztalatok azt mutatják, hogy a csak magyar nyelvű tanítóhalmazon képzett modell (PULI-GPT-3SX) helyesírása sokkal jobb, mint a Llama-2 7B modellé. Ez vélhetően az alapmodellek tanítási adatai miatt van, hiszen a Llama-2 modellek döntően angol nyelvű szövegeken tanultak, míg a magyar modell szinte kizárólag magyar nyelvű korpuszon.

A helyesíráson túl a szövegek összefüggősége, érthetősége is jobbnak bizonyul a magyar modell esetében. A következőkben egy kreatív írást kértem mindkét modelltől az alábbiak szerint:

Rendszerprompt nem volt megadva.

Prompt: „### Question: Írj egy rövid cikket, amiben a pécsi tűzoltók kimentenek egy macskát egy mély kútból”

A PULI-GPT-3SX alapú modell a következő választ adta:

„### Answer: Egy macska kétségbeesetten próbált kiutat találni egy mély kútból, de nem tudott kijutni. Szerencsére néhány bátor pécsi tűzoltó azonnal a segítségére sietett, és sikerült kiszabadítaniuk a macskát. Miután biztonságba helyezték, a tűzoltóknak sikerült megtalálniuk a cica gazdáját is. Az állat biztonságban van, és hamarosan visszatérhet

otthonába. Köszönjük a pécsi tűzoltóknak az életmentő segítséget! #Pécs #tűzoltóság #macskamentés #mélykút #katasztrófavédelem #hősiesmunka #cicamentés #mentőakciók #vízvezeték-szerelők #Pécs #Tűzoltók #Katasztrófaelhárítás #Macskamentés #Kútúrás #Mélykút”

A Llama-2 7B alapú modell a következő választ adta:

„### Answer: Őszintén meglepették, amikor a pécsi tűzoltóságnak feladatot adtak egy macska mentésére. A macska egy mély kútban rekedt, és sok órája tartózkodott ott. A tűzoltók nagyon óvatosan léptek be a kúthoz, hogy megtalálják a macskát, de végül sikerült elérniük. A macska épségben volt, és hamarosan új otthonra találták. Ez az eset bizonyítja, hogy a tűzoltók mindig készségesen segítenek másoknak, még ha veszélyes helyzetben is.

A macska mentése inspiráló történet volt, amely arra emlékezteti bennünket, hogy soha ne add fel,”

Mindkét modell megértette a feladatot és elkészítették válaszaikat. A magyar modell végig kontextusban maradt és a cikk végére hashtag-eket is lehelyezett. A hashtag-ek között már volt nem megfelelő, például a vízvezeték-szerelők és a kútúrás. A Llama-2 7B alapú modell válasza érthető ugyan és a témához kapcsolódó, de helyesírási hibáktól nem mentes, az utolsó mondat már nem szorosan kapcsolódik a feladathoz, tehát kiesett a kontextusból.

Az általános kérdésekre a modellek érthető, többnyire megfelelő válaszokat tudnak generálni, azonban a modellépítés célja a tűzoltó szakmai ismeretek elsajátítása és egy szakértői rendszer létrehozása.

A szakmai kérdésekre és az azokra adott válaszokból nézzünk egy példát:

Rendszerprompt nem volt megadva.

Prompt: „### Question: Milyen dokumentumokat kell beadnom egy tűzjelző berendezés engedélyezéséhez?”

A PULI-GPT-3SX alapú modell a következő választ adta:

Answer: Egy tűzjelző berendezés engedélyezéséhez be kell nyújtania a szükséges iratokat, például az üzemeltetési engedélyt vagy a tűzvédelmi megfelelőségi tanúsítványt. Ezenkívül be kell mutatnia a berendezés létesítési engedélyét és/vagy használatbavételi engedélyét is. Az engedélyezési dokumentációnak tartalmaznia kell a berendezés műszaki leírását, valamint a berendezéshez kapcsolódó dokumentációt, például a telepítési rajzokat, a kapcsolási rajzot, a villamos bekötéseket stb. Végül be kell adnia a berendezés üzembe helyezési jegyzőkönyvét, amely igazolja, hogy a berendezést megfelelően telepítették. Ha további információkra van szüksége, forduljon

hozzánk bizalommal! Szívesen segítünk Önnek bármilyen kérdésben.
A tűzjelző berendezések engedélyezésének folyamata általában több hónapot vesz igénybe, és jellemzően az illetékes hatóság végzi. Ez idő alatt a kérelmezőnek meg kell fizetnie a díjat, és bizonyítania kell, hogy rendelkezik a szükséges dokumentumokkal. A tűzjelző berendezéseket gyakran jóváhagyják, de bizonyos esetekben nem engedélyezik őket. Ilyen esetek közé tartozik a téves riasztás, a berendezés jogosulatlan használata, a berendezés módosítása vagy átalakítása, illetve, ha a berendezés nem felel meg a vonatkozó előírásoknak. Fontos megjegyezni, hogy a tűzjelző berendezés engedélyezése nem garantálja”

A Llama-2 7B alapú modell a következő választ adta:

„### Answer: 1. Tűzvédelmi tervezési dokumentáció (TtvD)
2. Kivitelezői nyilatkozat
3. A tűzjelző berendezés műszaki leírása, amely tartalmazza a jelzési zónák számát és az általuk védett helyiségeket.
4. Az épület/helyiség alaprajza, amely megmutatja a jelzési zónák határait.
5. Tervezői okirat vagy szakértői állásfoglalás, ha szükséges.
6. Helyszín-specifikus vizsgálati jegyzőkönyvek, ha szükséges.
7. Szimulációs elemzés, ha szükséges.
8. Építményszerkezeti terv, ha szükséges.”

Mindkét válasz értelmes, a laikus és a hozzáértő egyaránt megérti. Kiindulva abból, hogy a kérdést szakmabeliek is feltehetik, mindenképpen pontosításokra és helyesbítésekre van szükség, a válaszoknak a vonatkozó jogszabály általi szóhasználattal szükséges megjelenniük.

Az első (PULI-GPT-3SX alapú modell) válasz elején lévő üzemeltetési engedély nem létezik és nem is kell benyújtani. A második mondat sem fedi a valóságos jogszabályi környezetet, mivel létesítési engedélyezéshez sem létesítési, sem használatbavételi engedélyt nem kell benyújtani. A 3. mondat helytálló, azonban nem pontosak a megnevezések. Az üzembe helyezési jegyzőkönyvet használatbavételi engedélyezés során kell benyújtani. Az engedélyezés időtartamát tekintve az „általában több hónapot vesz igénybe” helyett a „néha/olykor több hónapot vesz igénybe” jobban fedi a valóságot, valamint „a kérelmezőnek meg kell fizetnie a díjat” nem helytálló megállapítás, nincs díjfizetési kötelezettség. A következő mondatban szereplő jóváhagyás elfogadható engedélyezés, de szakmai körökben félreértésre adhat okot, mivel a jóváhagyás külön hatósági eljárás. A következő mondatban felhozott példák közül „a téves riasztás, a berendezés jogosulatlan használata” nem megfelelőek. Az utolsó mondat nem befejezett.

A második (Llama-2 7B alapú modell) válasz csak felsorolást tartalmaz, de nem mindent azok közül, amit a vonatkozó jogszabály előír, tehát nem tekinthető

teljesnek. Ugyan szakmainak tűnik a szóhasználat, de nem jogszabályszerű, amelyet a tervezők, hozzáértők gyorsan észrevesznek.

Összességben az első válasz közérthetőbb, laikusok is könnyen értelmezik, azonban téves információkat is ad. Ezzel szemben a második, tömörebb válasz közelebb áll a jogszabályi követelményekhez, még úgy is, hogy néhol nem szakmai a szóhasználata, ezért magam részéről azt tartom jobbnak.

A modelleket vizsgálva megállapítható, hogy mindkettő többnyire képes megérteni a kérdéseket és szakmainak tűnő válaszokat képes generálni. Ezek nagyrészt kapcsolódnak a kérdéshez és szakmailag relevánsak, azonban időnként, főleg a hosszabb válaszok kapcsán kezdenek eltérni a témától és nem releváns dolgokat is belekevernek a válaszbba. Amikor a modellnek nincs pontos információja, alaptudása a kérdéssel kapcsolatban, akkor hajlamos hallucinálni, tehát nagyon magabiztosan közöl téves információkat.

Összefoglalás, következtetés

A modellek kiválasztása és finomhangolása a rendelkezésemre álló technikai háttér figyelembevételével történt meg. Ez a háttér rendkívül szerény hardveralapokon nyugszik, ahhoz képest, hogy a nyelvi modellezéshez milyen számítási háttér szükséges, például az alapmodellek tekintetében. Az eredmények számbavétele során megállapíthatjuk, hogy komolyabb adatközponti kapacitásokkal a bemutatottnál komolyabb eredmények érhetőek el, azonban a bemutatott eredmények is figyelemre méltóak.

A felmerülő problémák áttekintése

Elavuló tudás: a nyelvi modellek nem tudnak valós idejű vagy friss adatokat szolgáltatni, mivel kizárólag a számukra biztosított képzési adatokra támaszkodnak, ez egy szakértői rendszernél probléma lehet.

Interakciók hiánya: a nyelvi modellek nem hajthatnak végre műveleteket vagy nem léphetnek kapcsolatba külső rendszerekkel, ami korlátozza azok funkcionalitását. Például nem indíthatnak keresést az interneten további információkért, nem tudnak valós időben lekérdezni adatbázisokat, és nem használhatnak számológépet bonyolultabb matematikai számítások elvégzéséhez.

Kontextus és további információk hiánya a nyelvi modellnek nehézséget okozhat abban, hogy megértse és beépítse a korábbi promptokból vagy beszélgetésekből származó kontextust. Előfordulhat, hogy nem emlékeznek a korábban említett részletekre, vagy nem adnak további releváns információkat a megadott

felszólításon túl. Ezek elsősorban a kontextusmérettel vannak összefüggésben. A modell kizárólag a kontextusban átvett tokenekre tud támaszkodni a válasz generálása közben.

Hallucinációk: a nyelvi modellek sok általános tudással rendelkeznek a világról, ami a súlyukban rejlik. Előfordulhat azonban, hogy nem értenek kellőképpen bizonyos témákat és olyan válaszokat generálnak, amelyek nem tényszerűek vagy koherensek. Például olyan információkat állíthatnak elő, amelyek nem léteznek vagy pontatlan adatokat szolgáltatnak. Ezek a hibák a modell mértének növelésével és beállításokkal ugyan csökkenthetők, de teljesen nem lehet eltüntetni azokat.

Elfogultság: Attól függően, hogy milyen adatok alapján képezték ki a nyelvi modelleket, azok elfogultságokat mutathatnak, amelyek lehetnek vallási, ideológiai, politikai és egyéb természetűek.

Eredmények áttekintése

A kitűzött céloom egy olyan szakértői rendszer létrehozása volt, amely tűzvédelmi, katasztrófavédelmi szakkérdésekben képes releváns válaszokat adni. A nyelvi modellek, amelyeket bemutattam, csak korlátozottan képesek ezt a feladatot ellátni, hiszen viszonylag kis méretűek és a tanítóadatok mennyisége és minősége sem volt elegendően változatos. Azonban az a tény, hogy egy teljesen független, saját, védett környezetben futó nyelvi modellt tudtam készíteni, amely versenyképes lehet a piacon ma megtalálható modellekkel, mindenképpen inspiráló és előremutató eredmény.

A valódi versenyképes rendszer létrehozásához azonban olyan professzionális szintű technikai háttér szükséges, amely a kutatásban számomra nem volt elérhető.

Nagyobb modellmérettel, legalább 8 bites kvantálással vagy kvantálás nélküli modellek felhasználásával jobb eredményeket lehet elérni. A tanítóadatok gondos kiválasztásával, lehetőség szerint szakemberek által készített kérdés-válasz párosokkal a szintetikus adatokhoz képest sokkal jobb eredményeket lehet elérni (Zhou 2023).

A nyelvi modell képzésénél használt kontextusméret is meghatározó a képzést illetően, amit szintén a technikai adottságaimat tekintve kis méretre kellett beállítanom. A kisebb kontextusméretekben tanított modell nem képes átlátni a nagyobbakat, így kevésbé tud koncentrálni a hosszú kérdésekre, és a válaszadás során is kiesik a kontextusból a hosszú válaszok esetén.

A tűzvédelmi tesztek megoldása terén is javulást hozhat a nagyobb modellméret és több a tanítóadat alkalmazása.

Javaslatok

A nagyobb modellekről elmondható, hogy jobb teljesítményt nyújtanak, érdemes lehet 30 milliárd feletti paraméterszámú modelleket is kipróbálni. A Llamaz modell 70 milliárd paraméteres modellje még kezelhető valamilyen kvantálással, akár két 24 GB VRAM-ot tartalmazó GPU segítségével.

Több tanítóadat jobb közelítést jelent, amint azt az univerzális közelítési tételből is ismerjük (Hochreiter–Schmidhuber 1997). A jó minőségű tanítóadatok előállításuk kulcsfontosságú a jó minőségű kimenet előállításához. A tanítóadatok mennyiségét tovább kell növelni, akár több százszáz nagyságig, amely a modell általános minőségére jó hatással lenne.

A nagyméretű, előre betanított nyelvi modellek bizonyítottan tényszerű tudást tárolnak paramétereikben, és a legkorszerűbb eredményeket érik el, amikor finomhangolják azokat a későbbi NLP-feladatokra. Azonban – ahogy ezt láttuk a dolgozatban – a tudás, az ismeretek elérésére és pontos manipulálására való képességük még mindig korlátozott, ezért az úgynevezett tudásintenzív feladatokban – mint amilyen a tűzvédelmi szakértői rendszer kialakítása is – a teljesítményük elmarad a feladatspecifikus architektúrákétól. Ezért a kutatók olyan technológiát vezettek be, ahol a nyelvi modellt a válaszadásban más külső forrásból származó releváns információval segítik a válasz kialakításában. Ezeknek a RAG¹⁹ módszereknek a kutatása jelenleg is zajlik (Lewis 2020).

Ennél a módszertannál a rendelkezésünkre álló adatokat transzformálnunk kell egy látens térbe, ezt a folyamatot nevezzük beágyazásnak, majd egy kereső algoritmus segítségével keressük a releváns információkat. A nyelvi modell szerepe itt csupán annyi, hogy a kérdésből és a kikeresett releváns információkból választ kell generálnia.

Kutatásom során kipróbáltam ilyen módszereket is, és nagyon jó eredményeket kaptam, azonban ezeknek még további vizsgálata szükséges, amit a jövőben tervezek végrehajtani.

Irodalomjegyzék

- Baum, L. E. (1966) Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, 37(6). pp. 1554–1563. doi: [10.1214/aoms/1177699147](https://doi.org/10.1214/aoms/1177699147)
- Chorowski, J. B. (2015) Attention-based models for speech recognition. In *Neural Information Processing Systems*, pp. 577–585.
- Coles, C. (2023) <https://www.cyberhaven.com>.

<https://www.cyberhaven.com/blog/4-2-of-workers-have-pasted-company-data-into-chatgpt/>

[Letöltve: 2023.08.12.].

- Dettmers, T. P. (2023) QLoRA: Efficient Finetuning of Quantized LLMs. NeurIPS 2023 (oral) <https://arxiv.org/pdf/2305.14314.pdf>

[Letöltve: 2023.08.21.].

- Devlin, J.–Chang, M. W.–Lee, K.–Toutanova, K. (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

<https://arxiv.org/pdf/1810.04805.pdf>

[Letöltve: 2023.08.25.].

- Gordon, C. (2023) *Forbes*.

<https://www.forbes.com/sites/cindygordon/2023/02/02/chatgpt-is-the-fastest-growing-app-in-the-history-of-web-applications/?sh=6a83235f678c>

[Letöltve: 2023.07.04.].

- Hochreiter, S.–Schmidhuber, J. (1997) Long short-term memory. *Neural Computation*, pp. 1735–1780.

doi: [10.1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735)

- Hornik, K.–Stinchcombe, M.–Halbert, W. (1989) Multilayer Feedforward Networks are Universal Approximators. *Neural Networks*, Vol. 2. Pergamon Press., pp. 359–366.

doi: [10.1016/0893-6080\(89\)90020-8](https://doi.org/10.1016/0893-6080(89)90020-8)

- Hu, E.–Shen, Y.–Wallis, P.–Allen-Zhu, Z.–Li, Y.–Wang, S.–Wang, L.–Chen, W. (2021) LORA: Low-Rank adaptation of Large Language Models.

<https://arxiv.org/pdf/2106.09685.pdf>

[Letöltve: 2023.08.11.].

- Lewis, P. P. (2020) Retrieval-Augmented Generation for Knowledge-Intensive NLP tasks.

<https://arxiv.org/pdf/2005.11401.pdf>

[Letöltve: 2023.08.18.].

- Luccioni, A. S.–Viguiet, S.–Ligozat, A.-L. (2022) Estimating the carbon footprint of BLOOM, a 176B parameter language model.

<https://www.jmlr.org/papers/volume24/23-0069/23-0069.pdf> [Letöltve: 2023.08.14.].

- McGuire, N. (2018) How Accurate Is Google Translate in 2018?

<https://web.archive.org/web/20210125043932/https://www.argotrans.com/blog/accurate-google-translate-2018/> [Letöltve: 2023.08.02.].

- Mikolov, T. C. (2013) Efficient Estimation of Word Representations in Vector Space.

<https://arxiv.org/pdf/1301.3781> [Letöltve: 2023.08.05.].

- Radford, A. N. (2018) Improving language understanding by generative pre-training.

<https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language->

¹⁹ Retrieval Augmented Generation

- [unsupervised/language_understanding_paper.pdf](#)
[Letöltve: 2023.08.03.].
- Touvron, H.–Lavril, T.–Izacard, G.–Martinet, X.–Lachaux, M.-A.–Lacroix, T.–Rozière, B.–Goyal, N.–Hambro, E.–Azhar, F.–Rodriguez, A.–Joulin, A.–Grave, E.–Lample, G. (2023a) *LLaMA: Open and Efficient Foundation Language Models*.
<https://arxiv.org/abs/2302.13971>
[Letöltve: 2023.08.04.].
- Touvron, H. et al. (2023b) Llama 2: Open foundation and Fine-Tuned chat models.
<https://arxiv.org/pdf/2307.09288.pdf>
[Letöltve: 2023.08.03.].
- Vaswani, A. S. (2017) Attention Is All You Need. *Advances in Neural Information Processing Systems*, pp. 5998–6008.
- Wang, Y. K. (2022) *Self-Instruct: Aligning Language Models with Self-Generated Instructions*. doi: [10.48550/arxiv.2212.10560](https://doi.org/10.48550/arxiv.2212.10560) [Letöltve: 2023.08.12.].
- Weizenbaum, J. (1966) ELIZA – a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1). pp. 36–45.
- Yang, J. J. (2023) *Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond*.
<https://arxiv.org/abs/2304.13712>
[Letöltve: 2023.08.16.].
- Yang, Z. G. (2023) Jönnek a nagyok! BERT-Large, GPT-2 és GPT-3 nyelvmodellek magyar nyelvre. *XIX. Hungarian Computational Linguistics Conference*, pp. 247–262.
- Zhou, C. L. (2023) LIMA: Less is more for alignment.
<https://arxiv.org/abs/2305.11206>
[Letöltve: 2023.08.25.].