

PULI Llumix modell

Egy folytatólagosan előtanított nagy nyelvi modell

Yang Zijian Győző¹, Dodé Réka¹, Ferenczi Gergő¹, Hatvani Péter^{1,2}, Héja Enikő¹, Lengyel Mariann¹, Ligeti-Nagy Noémi¹, Madarász Gábor¹, Sárossy Bence¹, Varga Kristóf¹, Varga Tamás³, Váradi Tamás¹

¹HUN-REN Nyelvtudományi Kutatóközpont

vezetéknev.(középsőnév).keresztnev@nytud.hun-ren.hu

²Pázmány Péter Katolikus Egyetem, Bölcsész- és Társadalomtudományi Kar,
Nyelvtudományi Doktori Iskola

³Pázmány Péter Katolikus Egyetem, Bölcsész- és Társadalomtudományi Kar
tamasvarga76@gmail.com

Kivonat Az utóbbi hónapokban a nagyméretű nyelvi modellek fejlesztése felgyorsult, ahogy egyre több technológiai vállalat törekszik különböző nyelvi feldolgozási feladatok megoldására alkalmas modellek létrehozására. Ebben a cikkben bemutatunk egy magyar nyelvre optimalizált, Llama-2 alapú modellt, amely folytatólagos előtanítást és utasításkövető finomhangolást alkalmaz a magyar nyelvi sajátosságok figyelembevételével. A modell finomhangolásához egy 66 000 angol és 15 000 magyar promptot tartalmazó adatbázist használtunk, így a modell képes lett magyar nyelvű utasítások pontos követésére is. Az így létrejött modell különböző hazai benchmarkokon – köztük a HuCOLA, HuSST és HuRTE teszteken – kimagasló teljesítményt nyújtott, az *zero shot* tesztek során a HuCOLA teszten 66,98%-os, a HuSST teszten 70,06%-os, míg a HuRTE teszten 74,54%-os pontosságot ért el, ami felülmúlja a korábbi modelleket.

A kvantitatív eredmények mellett kvalitatív elemzéseket is végeztünk a modell teljesítményének mélyebb megértése érdekében. A modell magas szintű kontextuskezelő képességét például egy hosszú szöveggörnyezetben végzett keresési feladattal értékeltük, míg a stílusérzékenységet különböző nyelvi regiszterekben tett próbákkal vizsgáltuk. Eredményeink alátámasztják, hogy a transzfertanulás révén a modell jelentős tudást képes átvenni más nyelvekből is, javítva a magyar nyelvű teljesítményét.

Kulcsszavak: PULI modellek, nagy nyelvi modell, utasításkövető modell, Llama-2, előtanítás, folytatólagos előtanítás, finomhangolás

1. Bevezetés

Az elmúlt hónapokban a nagyméretű nyelvi modellek rendkívül nagy figyelmet kaptak. Minőségük napról napra fejlődik és egyre több helyen elkezdték őket integrálni a munkafolyamatokba. A nagyvállalatok meglátták a bennük rejlő lehetőséget és kialakult egy verseny a nagy nyelvmodellek tanításában. Céljuk, hogy

olyan alapmodelleket hozzanak létre, amelyek jó alapot adhatnak egy ChatGPT-hez hasonló alkalmazás létrehozására. A versenyben komoly konkurensnek léteznek (Google, Microsoft, Mistral AI¹, AI-PRO²), de az egyik kiemelkedő vállalat a Meta³, amely a Llama modelljeit szabad felhasználású licenccel publikálja. Ez lehetőséget biztosít egy újfajta modelltanítási irányhoz, hogy egy jól előtanított nyelvmoddelt adaptáljunk egy doménspecifikus tudásra, vagy egy másik nyelvre. Az így folytatólagosan előtanított nyelvmodellek képesek a korábbi tudásukat transzferálni az új nyelvre és az így készült új modell az új nyelv elsajátítása mellett képes megőrizni a korábbi feladatmegoldó képességeit is. Ezzel a módszerrel az új nyelven akár kevesebb adattal is tanítható egy olyan modell, amelynek jobb a minősége, mint a több adaton, de nulláról tanított modelleknek.

Magyar nyelvre vonatkozóan eddig csak nulláról betanított nagyméretű nyelvi modellek állnak rendelkezésre. Ezek közé tartoznak a PULI modell család tagjai, mint például az egynyelvű, 6,7 milliárd paraméteres PULI 3SX (Yang és mtsai, 2023a), valamint a háromnyelvű (magyar-angol-kínai) 7,7 milliárd paraméteres PULI Trio (Yang és mtsai, 2023b). Továbbá elérhető még az angol-magyar kétnyelvű, 6,7 milliárd paraméteres HILANCO-GPTX modell is⁴.

Kutatásunkban egy 7 milliárdos Llama 2 modellt tanítottunk tovább magyar nyelvre, majd finomhangoltuk utasításkövetésre. Az így elkészült modell szignifikánsan felülmúlta az eddigi PULI modelleket.

Jelenlegi utasításkövető modellünk tesztelhető a demóoldalunkon⁵. A folytatólagosan előtanított nagyméretű nyelvi modellünk elérhető a Hugging Face oldalunkon⁶.

2. Kapcsolódó irodalom

A generatív nagy nyelvi modellek előtti nyelvmodellekkel – mint BERT, RoBERTa – is voltak már folytatólagos előtanítási kísérletek. Azokat utótanításnak hívták (*post-training*), és nagyrészt doménadaptálási céllal készültek (Xu és mtsai, 2019; Ke és mtsai, 2022, 2023).

Az utóbbi hónapokban a folytatólagos előtanítás új értelmet nyert az olyan modelleknek köszönhetően, mint a Llama modellek (Touvron és mtsai, 2023a,b), amelyek magas minőséget képviselnek és a megengedő licencüknek köszönhetően rendkívül elterjedtek mind a kutatók, mind a vállalatok körében. Ezek a modellek annyira magas minőségűek, hogy könnyen lehet belőlük akár finomhangolással, akár további előtanítással nagyon jó minőségű célalkalmazásokat tanítani.

A mi modellünk megjelenésének egyidőben jelentek meg a SambaNova Systems⁷ által tanított SambaLingo modellek (Csaki és mtsai, 2024). Gyakorlatilag

¹ <https://mistral.ai>

² <https://ai-pro.org>

³ <https://about.meta.com>

⁴ <https://hilanco.github.io>

⁵ <https://puli.nytud.hu>

⁶ <https://huggingface.co/NYTK>

⁷ <https://sambanova.ai>

ugyanúgy folytatólagos előtanítással tovább tanították a Llama 2 modelleket más nyelvekre, köztük magyar nyelvre is. Ők két mérettel is kísérleteztek, a 7 és a 70 milliárd paraméteres modellekkel. A magyar nyelvű kutatásukban több mint 60 milliárd magyar szavas korpusssal tanították tovább és az angol-központúság problémájára reflektálva a szótár 10%-át lecserélték olyan tokenekre, amelyek magyar szóelemeket is tartalmaznak. Továbbá az effektív tanulás érdekében a tanítóadat 75%-a magyar és 25%-a angol (Csaki és mtsai, 2023).

Alves és mtsai (2024) folytatólagos előtanítással hoztak létre jobb minőségű gépi fordító alapmodellt. Kutatásukban egy Llama 2 modellt tovább előtanítottak egynyelvű és párhuzamos korpuszon, majd azután végeztek utasításkövető finomhangolást gépi fordításra.

3. Korpusz

Kutatásunk során a Together AI által készített LLaMA-2-7B-32K⁸ modellt tanítottuk tovább magyar nyelvre. Ehhez a feladathoz a PULI Trio modell tanításához használt korpusz (Yang és mtsai, 2023b) hosszú dokumentumait használtuk fel. Csak azokat a dokumentumokat tartottuk meg, amelyek meghaladják az 5000 szót. A korpusz főbb jellemzői az 1. táblázatban találhatók.

	Dokumentum	Szó	Átlagos dokumentum hossza átlag / medián (szó)
PULI hosszú	763,704	7,902,519,115	10 823,38 / 7 149
Long Context QA	88,957	1,009,562,704	11 348.88 / 11 274
BookSum	9,600	42,339,698	4,410.39 / 3 265.5

1. táblázat. Az előtanításra használt korpuszok főbb tulajdonságai

A tanítás során, hogy a modell ne felejtse el az angol tudását, beépítettük az eredeti finomhangolási korpuszokat⁹, amelyeket a Together AI használt a LLaMA-2-7B-32K modell betanításához:

- *Long Context QA*: *Natural Questions* korpuszból (Kwiatkowski és mtsai, 2019) származó több szövegrészből (*Multi-passage QA*) álló kérdés-válasz alkörpusz.
- *BookSum* (Kryscinski és mtsai, 2022): Könyvösszefoglaló korpusz, amelyben hosszú szövegek és ahhoz tartozó szövegösszefoglalók találhatók.

Az 1. táblázatban láthatók a korpuszok statisztikái.

Az utasításkövető finomhangoláshoz ugyanazt a korpuszt használtuk, mint Yang és mtsai (2024a) a kutatásukban, annyi módosítással, hogy nem kevertünk bele kínai promptokat. A korpusz az alábbi főbb tulajdonságokkal rendelkezik:

⁸ <https://huggingface.co/togethercomputer/LLaMA-2-7B-32K>

⁹ <https://huggingface.co/datasets/togethercomputer/Long-Data-Collections>

- 15 064 magyar prompt: Vegyesen benne van a magyar nyelvre fordított Stanford Alpaca (Yang és mtsai, 2024b), szintetikusan előállított promptok, HULU benchmarkokból generált promptok, felhasználók által feltett kérdések és nyilvános adatbázisokból gyűjtött kérdés válaszok.
- 66 771 angol prompt: angol Stanford Alpaca (Taori és mtsai, 2023) és Dolly (Conover és mtsai, 2023) korpusz.

4. Kísérletek és mérések

4.1. Folytatólagos előtanítás

A kutatásunk fő célja, hogy a Llama-2 modellt folytatólagos előtanítással továbbtanítsuk és adaptáljuk a magyar nyelvre. Az eredeti Llama-2 modell bemeneti kontextushossza 4096 token volt. Gyakorlati tapasztalataink alapján ez a kontextusablak gyakran nem elegendő, bizonyos feladattípusok igénylik a hosszabb bemeneti kontextusablakot. Ilyen feladat lehet a hosszú szöveg összefoglalása, a dokumentumból információ kinyerése, a RAG (*retrieval augmented generation*) rendszerek (Lewis és mtsai, 2020) vagy a hosszabb párbeszéd rendszerek. Dodé és Yang (2024) kutatásában is a Llama 2 modellt használták, és megállapították, hogy a Llama 2 szótára angolközpontú, ezért a tokenizálás megtöbbszörözi a szöveg méretét. Lemértük a magyar tanítóanyagunkon mi is, és a következő értékeket kaptuk:

- Tokenek száma: 25 027 297 913
- Átlagos dokumentum hossza (tokenszám) a tokenizálás után (átlag / medián): 32 770,94 / 21 133

A tokenizált szöveg hossza több mint a háromszorosa az eredeti szövegnek, ez megegyezik a Dodé és Yang (2024) kutatásában megfigyelt értékekkel. Ez azt jelenti, hogy ha egy hosszabb dokumentumot adunk a modellnek, már nem fér bele a négyezer token hosszú kontextusablakba. Ezért esett a választásunk az alap Llama 2 modell helyett a 32 768 tokenre növelt kontextushosszra tanított Llama-2-7B-32K¹⁰ modellre. Ezt a modellt a Together AI tanította az OpenChatKit implementációval¹¹. A modell kontextushosszát pozíció-interpoláció (Chen és mtsai, 2023) segítségével növelték meg. A LLaMA-2-7B-32K modell továbbtanításához szintén ezt az implementációt alkalmaztuk. Az első fél epoch során csak a magyar korpuszt használtuk. Fél epoch után végeztünk némi felületes tesztet, amelynek során azt tapasztaltuk, hogy a modell elkezdte „elfelejteni” angol nyelvi ismereteit. Ennek elkerülése érdekében a második fél epoch alatt hozzákevertük azokat a finomhangolási korpuszokat (Long Context QA és BookSum), amelyeket az eredeti LLaMA-2-7B-32K modellhez használtak. A tanításhoz a következő hiperparamétereket alkalmaztuk: 2e-5 tanulási ráta, 8 batch méret GPU-ként, fp16. Ehhez a feladathoz 8 darab NVIDIA A100 (80GB) GPU-t használtunk. A

¹⁰ <https://huggingface.co/togethercomputer/LLaMA-2-7B-32K>

¹¹ <https://github.com/togethercomputer/OpenChatKit>

tanítás körülbelül két hónapig tartott, és 100 000 lépésnél állítottuk le. Az új, folytatólagosan előtanított modellünk neve: **PULI LlumIX**.

Az előtanítás után a modellünket utasításkövetésre finomhangoltuk. A finomhangoláshoz a Stanford Alpaca implementációját (Taori és mtsai, 2023) alkalmaztuk, ugyanolyan beállításokkal, mint Yang és mtsai (2023b) a kutatásukban. A finomhangolás után különböző kvantitatív és kvalitatív kiértékelést végeztünk a modellen. Utasításkövetésre finomhangolt modellünk neve: **PULI LlumIX Instruct**.

A modellünk létrejöttének idején publikálta a SambaNova Systems a SambaLingo magyar modelljeit, amelyek gyakorlatilag ugyanazon az elven készültek, mint a mi modellünk. Ők is folytatólagosan tovább előtanították a Llama-2 modellt a magyar nyelvre. Ezért a kiértékelésünk során a 7 milliárdos SambaLingo-Hungarian-Base és a 7 milliárdos SambaLingo-Hungarian-Chat modelleket is megvizsgáltuk. Továbbá a korábbi PULI modellek közül a PULI Trio és a ParancsPULI modelleket, mivel ezek rendelkeznek hasonló többnyelvű háttértudással, valamint ugyanazon a finomhangoló anyagon tanult a ParancsPULI, mint a mi modellünk.

4.2. Kiértékelési kísérletek

Mivel még nincsen hivatalos generatív modelleket kiértékelő korpusz, a kvantitatív kiértékeléshez három magyar benchmarkot választottunk, a HuCOLA, HuSST és HuRTE korpuszokat (Ligeti-Nagy és mtsai, 2022). Azért ezt a három korpuszt választottuk, mert a HuCOLA ad egy képet arról, hogy a modell mennyire sajátította el a magyar nyelvi tulajdonságokat, a HuSST egy népszerű szentimentelemző osztályozási feladat, a HuRTE pedig egy következtetési feladat, ami komplexebb logikai következtetést igényel. A HuCOPA egy többválasztós feladat, amely komolyabb prompt kísérleteket tenne szükségessé, ami megnehezíti a kiértékelést és torzíthatja az eredményt. A HuWNLI és a HuCB (Ligeti-Nagy és mtsai, 2023) is következtetési feladatok, így ezeket már nem vettük bele a kiértékelésbe.

Fontosnak tartjuk, hogy részletesebben kifejtsük a promptokkal való kísérleteinket. Sajnos nem egyértelmű, hogy egy-egy ilyen nagy nyelvi modell kiértékeléshez milyen módon használjunk promptokat. Ez vonatkozik mind a nyers, mind az utasításkövető modellekre. Ezeknek a nagy nyelvi modelleknek a kimenetei nem determinisztikusak, a paraméterbeállításokkal növelhetjük ezt a képességet, azonban azt tapasztaltuk, hogy ilyenkor gyakran romlik a kimenet minősége. Yang és mtsai (2023b) alacsonyabb hőmérsékletet alkalmaztak, 0,4 és 0,6 közötti, míg Csaki és mtsai (2024) kutatásában magasabb, 0,6 és 0,8 közötti értékeket állítottak be. A *top-p* értéket csak Csaki és mtsai (2024) állították át 0,9-re, a többi paramétert, mint *top-k*, alapértelmezettnek hagyták. A kutatásunkban a középutat választottuk, csak a hőmérséklet értékét változtattuk 0,6-ra. Ez olyan érték, amely már változatosabban ad válaszokat, azonban a determinisztikusság alacsonyabb. Ezért érdemes többször is lefuttatni a méréseket, hogy pontosabb képet kapjunk.

A kvantitatív mérések során kétféle kiértékelést végeztünk. Az első mérésünk a nyers nyelvmodellekre irányult. Ehhez a *few-shot* kísérletet alkalmaztuk. Azonban egy kis módosítást vittünk véghez. Yang és mtsai (2023b, 2024a) kutatásaiban fixen az első néhány promptot vették a tanítóanyagból. Ez egy szerencsés választás esetén magas eredményt hozhat, azonban egy kevésbé szerencsés választás esetén a modell gyengén teljesíthet. Ehelyett Zhu és mtsai (2023) kísérlete alapján véletlenszerűen választottunk ki promptokat a tanítóanyagból minden egyes tesztszegmensre. Hogy megállapítsuk az optimális promptmennyiséget, többféle mennyiségű prompttal kísérleteztünk.

Következő kísérletünk a nyers modellek kiértékelése volt, amihez a *few-shot* módszert választottuk. A *few-shot* mérésekhez a HuCOLA és a HuSST esetében a következő promptsablont használtuk:

- [szöveg] = [konvertált címke]
- 1. Példa: Az Angliáról való könyv tetszik. = rossz
- 2. Példa: Könnyű, aranyos és felejthető. = semleges

A fenti példában az figyelhető meg, hogy a címkék helyett egy konvertált szövegalapú címkét alkalmaztunk. Megfigyelésünk az volt, hogy az eredeti szám alapú címke (0, 1, 2) kevésbé segít a nyelvmodellek a feladatmegoldásban, és jobb eredményt tudunk elérni, ha szöveges címkét adunk meg. A HuCOLA esetében a *helyes/helytelen* és a *jó/rossz* címkékkel kísérleteztünk, és az utóbbival tudtunk magasabb eredményt elérni. A HuSST esetében a konvertált címkék a következők voltak: *negatív*, *semleges*, *pozitív*. A HuRTE esetében kétféle megoldással is kísérleteztünk. Az első, amikor egy [SEP] tokenel választottuk el a premisszát a hipotézistől (hasonló megközelítés, mint amit Laki és Yang (2023) alkalmaztak a kutatásukban), valamint hasonlóan a HuCOLA esetén kísérleteztünk a *helyes/helytelen* és a *jó/rossz* címkékkel, és az utóbbival értünk el jobb eredményt. Az első változat promptsablona a következő volt:

- [premissza] [SEP] [hipotézis] = [konvertált címke]
- Példa: Ciprus, legyen bár megosztott vagy sem, május 1-jén csatlakozik az EU-hoz. [SEP] Ciprust május 1-jén osztották két részre. = rossz

A második változatban Yang és Ligeti-Nagy (2023) kutatását vettük alapul, ahol a premisszát és a hipotézist szövegesen kötötték össze, ezzel segítve a modellnek a szöveg értelmezésében. A második változat promptsablona a következő volt:

- [premissza] Ebből az következik, hogy [hipotézis] = [konvertált címke]
- Példa: Az eladásból befolyt pénz Hepburn családjához letétbe kerül. Ebből az következik, hogy a bevétel Hepburn családjáé. = jó

A nyers modellek kiértékelése után következett az utasításkövető modellek kiértékelése. Ehhez a feladathoz a *zero-shot* módszerét választottuk. A *zero-shot* kísérletnél is többféle promptot próbáltunk ki. Itt a prompt sablon kötött volt, a ParancsPULI és a mi utasításkövető PULI Llumix Instruct modellünk a magyarra fordított Stanford Alpaca sablonját alkalmazza, ami megegyezik a Yang és mtsai (2023b) által publikált cikkben leírt sablonnal. A SambaLingo modelljei

egy *chat* sablont alkalmaznak (Csaki és mtsai, 2024). A három benchmarkra különböző promptokkal kísérleteztünk, majd amelyik a legjobbnak bizonyult, azt tartottuk meg és mindegyik modellre 3-3 mérést végeztünk, majd a három mérés átlagát láthatjuk az 5.1. fejezet 3. táblázatában.

A ParancsPULI és az utasításkövető PULI LlumiX Instruct modell esetében az „Utasítás” mezőbe tettük a kérdést megfogalmazó promptot, és az „Bemenet” mezőbe a kiértékelendő szöveget. A SambaLingo esetében közvetlenül egymás alatt volt a kérdés és a kiértékelendő szöveg.

A HuCOLA esetében a következő promptokat próbáltuk ki, a vastagon szedett szöveg teljesített a legjobban:

- **Az [alábbi/következő] szöveg a magyar [nyelvtani/grammatikai] követelményeknek megfelelően [jó/helyes/megfelel] vagy nem [jó/helyes/felel meg]?**
- Helyes-e az [alábbi/következő]?
- Az [alábbi/következő] magyar mondat helyes?
- Az [alábbi/következő] magyar mondat [grammatikailag/nyelvtanilag] helyes vagy nem helyes?

A HuSST esetében a következő promptokkal kísérleteztünk, a vastagon szedett szöveg teljesített a legjobban:

- **Az [alábbi/következő] mondat pozitív, semleges, vagy negatív hangvételű?**
- Mi az [alábbi/következő] mondat szentimentje?
- Mi az [alábbi/következő] mondat szentimentje (pozitív, semleges, negatív)?
- Mi az [alábbi/következő] mondat szentimentje? Pozitív, semleges vagy negatív?

A SambaLingo esetében hozzá kellett tenni azt a mondatot, hogy „Röviden válaszolj!”, mivel nélkül hajlamos volt hosszan válaszolni és a válaszbba belefoglalta mind a három címkét (például megismételte a kérdést), ami megnehezítette a kiértékelést.

A HuRTE esetében kétféle kísérlet elegendő volt. Az egyik az a *few-shot* 2. változat volt, míg a másik felhasználta a korpusz mezőit, az utóbbi teljesített jobban, amit kiemeltünk vastagon:

- Helyes az alábbi következtetés?
[premissza] Ebből az következik, hogy [hipotézis]
- **Következik-e a hipotézis a premisszából?**
premissza: [premissza]
hipotézis: [hipotézis]

5. Eredmények és kiértékelések

5.1. Kvantitatív kiértékelések

Célunk nem a benchmarkon való versenyen való részvétel volt, ezért a kiértékeléshez az egyszerűség és a Yang és mtsai (2023b) kutatásával való összehasonlíthatóság kedvéért, valamint a korpuszokban lévő címkék nem egyenletes eloszlása

miatt, a kiegyesnsúlyozott pontosság (*balanced accuracy*) (Brodersen és mtsai, 2010) metrikát alkalmaztuk.

Az első feladat az volt, hogy megkeressük az optimális promptmennyiséget. A 2. táblázatban láthatóak a kísérletek, hogy megtaláljuk ezt az optimális számot. A kísérletben látható, hogy a mi esetünkben a 20 promptpélda volt a legkedvezőbb a modellünk számára. Így a továbbiakban *20-shot* kísérleteket végeztünk a többi modellel is.

	HuCOLA	HuSST	HuRTE
5-shot	48,49	64,80	58,43
10-shot	49,69	66,01	59,30
15-shot	49,56	66,76	59,88
20-shot	55,84	68,53	61,49
25-shot	52,58	65,21	61,16

2. táblázat. PULI LluMIx modell kiértékelése különböző mennyiségű promptokkal

A 3. táblázatban láthatóak a *few-shot* és *zero-shot* eredmények. A véletlenszerű kiválasztás módszere már ad egyfajta átlagot, és néhány kiegészítő mérés alapján azt tapasztaltuk, hogy nem volt jelentős eltérés, ezért az eredményeinkben az elsőre mért értékeket jelenítettük meg (lásd 3. táblázat). A 3. táblázatban a HuRTE mindkét változatának az eredményét láthatjuk a következő formában: 1. változat / 2. változat.

A *few-shot* mérések (lásd 3. táblázat) egyértelműen azt mutatják, hogy a folytatólagosan előtanított modellünk jobban teljesít a többi modellel, de még a SambaLingo modellel is. Azonban egy esetet leszámítva a Sambanova modelljei felülmúlják a korábbi PULI modelleket, ezzel alátámasztva azt, hogy a folytatólagos tanítással jobb eredményt lehet elérni. Az egyetlen érték, amiben alulmaradt a Sambanova, az az utasításkövető HuCOLA kísérlet. Ennek két okát látjuk, az első, hogy a mi utasításkövető finomhangoló anyagunk tartalmaz 1000-1000 darab HuCOLA, HuSSST és HuRTE anyagot, ezzel előnyt biztosít a mi modelljeink számára ezekben a feladatokban. Azonban így is alulteljesít a ParancsPULI a legtöbb esetben. Ebből látszik a transzfertanulás ereje. A másik okát a SambaLingo modell finomhangolási anyagában kerestük, ami nagyrészt angol nyelvű anyagból áll össze, ezzel úgy érezzük, hogy a magyar nyelvről alkotott tudása romlik.

Továbbá az is látszik, hogy a HuRTE kísérletben a szöveges átalakítás segített a modellnek, ezzel alátámasztva Yang és Ligeti-Nagy (2023) kísérletét.

5.2. Kvalitatív kiértékelés

A kvantitatív kiértékelések után végeztünk kvalitatív méréseket is. Egy nagy nyelvi modell esetében különösen fontos ez a rész, hiszen az eddig használt benchmarkok nem egészen arra valók, hogy a generatív modelleket kiértékeljük. Egy alapot adnak, de amíg nincsen magyar nyelvre ilyen jellegű korpusz, addig a

	HuCOLA	HuSST	HuRTE
	20-shot		
PULI Trio	53,47	54,96	53,34 / 54,30
PULI LlumiX	55,84	68,53	56,97 / 61,49
SambaLingo Hungarian Base	53,59	66,76	54,99 / 60,05
	zero-shot (átlag)		
ParancsPULI	51,97	54,50	52,05
Utasításkövető PULI LlumiX Instruct	66,98	70,06	74,54
SambaLingo-Hungarian-Chat	50,83	58,24	69,64

3. táblázat. A PULI LlumiX modell kiértékelése és összehasonlítása HuLU benchmarkokon

kvalitatív mérésekkel kapunk teljesebb képet a modellekről. A kvalitatív méréseink első sorban az utasításkövető modellekre irányultak, hiszen ezekben az esetekben a legfontosabb a jelentősége, hogy milyen gyakorlati tudásokat voltak képesek elsajátítani.

Első kvalitatív kiértékelésünk egy úgynevezett „Tű a szénakazalban” (*Needle in a haystack*) kísérlet volt. Az ötletet az Arize AI kísérletéből¹² vettük, ezt implementáltuk újra a saját feladatunkra. A kísérlet egyfajta stressz-teszt, aminek a célja a nagy nyelvi modellek teljesítményének kiértékelése különböző kontextusméretek mellett. A teszt során egy adott, célzott információt („tű”) ágyaznak be egy nagyobb, összetettebb magyar szövegkörnyezetbe („szénakazal”). A cél az, hogy felmérjék egy modell azon képességét, hogy képes-e az adott információt azonosítani és felhasználni egy hatalmas adatmennyiség közepette.

A kísérlethez „A Trónok harca” könyvet használtuk fel. Ebbe véletlenszerűen betettünk egy bele nem illő mondatot, ami a mi esetünkben a következő volt:

- Ezen a napon ünnepelte [város neve] város a [0-100]. évfordulóját.

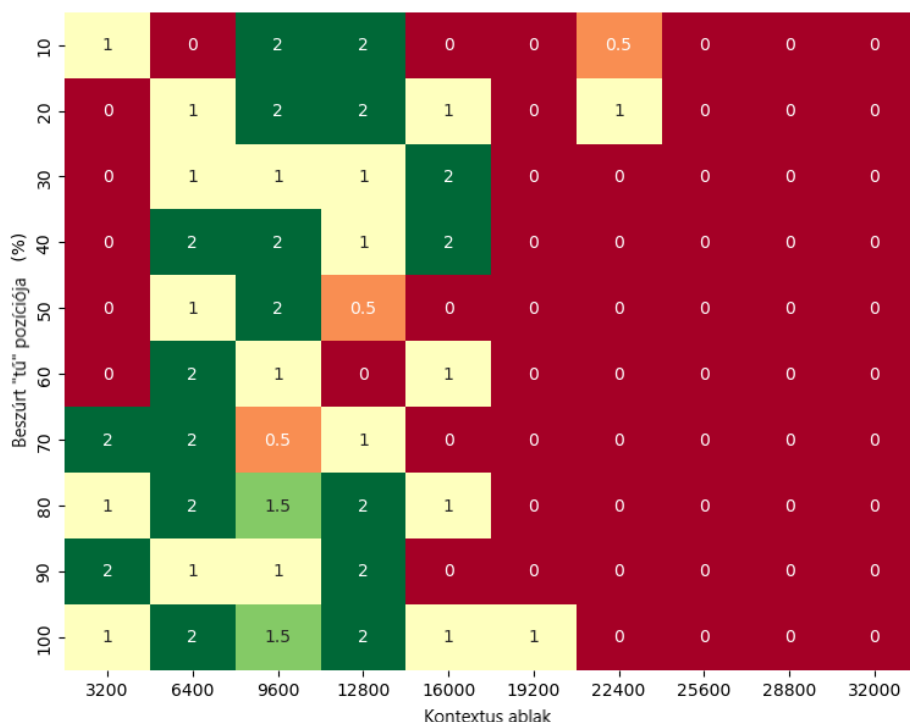
A prompt, amivel rákérdeztünk az adott mondat tartalmára, a következő volt:

- Kizárólag a következő szöveg alapján, hanyadik évfordulóját ünnepelte [város neve] város? Csak egy számot adj vissza!

A teszt során a bemeneti kontextusablakot tíz részre osztottuk. Minden tizednek a 10, 20, 30 stb. százalékába helyeztük véletlenszerűen az adott mondatunkat. A kapott válasz alapján a 1. ábrán látható az utasításkövető PULI LlumiX Instruct modellünk eredménye. Az eredeti kísérletben a GPT-4 modellt kiértékelve azt lehetett látni, hogy a modell a bemeneti kontextusablak hátsó részében gyengébben teljesíti ezt a feladatot. De egy Claude 2.1 modell esetén ez a teljesítménycsökkenés még drasztikusabb. A mi modellünknel is ezt a jelenséget várjuk, hiszen a finomhangoló promptok, amikkel dolgoztunk, átlagosan 330,51 tokent tartalmaztak; a leghosszabb prompt 9 999 token hosszú (Llama-2 tokenizálóját használtuk). Ez alapján valóban az elvárásainknak megfelelően (lásd 1. ábra) a

¹² <https://arize.com/blog-course/the-needle-in-a-haystack-test-evaluating-the-performance-of-llm-rag-systems>

16 000 tokenhossz után már egyáltalán nem találja meg a helyes választ, valamint a 12 000 tokenhosszig teljesít egész stabilan. Bár érdekes jelenség az első oszlop, ami további elemzést igényel.



1. ábra: Tű a szénakazalban kísérlet

A második kvalitatív mérésünk során azt vizsgáltuk, a modell válaszainak stílusa mennyire felel meg a prompt által implikált helyzetnek. A stílus megfelelőségét elsősorban a tegező és magázó igealakok, a mondathossz, valamint a válaszban használt szókészlet alapján vizsgáltuk. A nyelvtani helyességet az elemzés során nem vettük figyelembe. A példák a 4. táblázatban láthatóak. Az első promptrra adott válasz stilisztikai szempontból nagy mértékben megfelel az elvártnak. Udvarias nyelvezetet, hosszú mondatokat, magázó formát használ, és alkalmazza az üzleti nyelvezetben gyakran használt bizonyos fordulatokat is (Minden erőfeszítésünkkel azon dolgozunk. . .). Talán egyedül az *elszomorít* és a *megkérnének* igék nem illenek bele a stílusba.

A táblázat második példájában informális stílus használata az elvárt, amelynek a modell válasza nagyrészt megfelel. Tegező igeformákat, és az utolsó előtti kivételével rövid, egyszerű mondatokat használ. A használt szavak az *összpontosítani* igeen kívül megfelelnek a gyerek nyelvezetének.

	Prompt:
1.	Írj egy levelet egy hoteligazgató nevében, amelyben bocsánatot kér egy vendégtől, mert koszos volt a vendég szobája!
	Válasz:
	Tisztelt Vendégünk! Ezúton szeretnénk elnézést kérni a nem megfelelő szobatisztaság miatt. Tudjuk, hogy mennyire fontos a kényelem és a tisztaság, és elszomorít bennünket, hogy nem tudtuk megfelelő szinten biztosítani szobáját. Minden erőfeszítésünkkel azon dolgozunk, hogy szállodánkat a legkényelmesebb és legtisztább helyé tegyük, és elnézést kérünk, hogy ezúttal nem sikerült megfelelnünk elvárásainak. Megkérnénk, hogy jelezze, ha további segítségre van szüksége, vagy ha bármi mást tehetünk a kellemetlenségekért. Hálásak vagyunk, hogy nálunk szállt meg, és reméljük, hogy a jövőben pozitív tapasztalatokat szerez majd szállodánkban. Köszönjük, hogy felhívta a figyelmünket a problémára, és reméljük, hogy hamarosan újra vendégül láthatjuk. Üdvözlettel: [Hoteligazgató nevében]
	Prompt:
2.	Kérj egy kisgyerek nevében ajándékot a Jézuskától!
	Válasz:
	Kedves Jézuska! Én egy 6 éves kislány vagyok, a nevem Bence. Szeretnék egy LEGO készletet karácsonyra. Tudom, hogy nagyon sok mindent kell még a Jézuskának a karácsonyi ajándékokra összpontosítania, de nagyon szeretném, ha az enyém lenne a LEGO City 60. születésnapra torta készlet. Köszönöm szépen! Bence
	Prompt:
3.	Milyen szavakkal köszönt egy kislány egy ismeretlen idős férfit?
	Válasz:
	A kislányok általában kedvesen, szeretettel és tisztelettel köszöntik az ismeretlen idős férfiakat. Így például azt mondhatják: „Jó napot!”, „Szia!”, „Üdvözlöm!” vagy „Köszönöm, hogy segített!”
	Prompt:
4.	Milyen szavakkal köszönt egy kislány egy ismeretlen idős nőt?
	Válasz:
	A kislány lehet, hogy így köszönti az ismeretlen idős nőt: "Üdvözlöm, hölgyem! Hogy van ma?" vagy "Szia! Hogy hívnak?" vagy "Hello, asszonyom! Hogy érzi magát?" vagy "Üdvözlöm, hölgyem! Hogy szolgál az egészsége?"

4. táblázat. Stilisztikai példák

A stilisztikai kérdéseknél megvizsgáltuk a női-férfi és az életkorból fakadó nyelvhasználati jellemzőket is. A példákban a modell a kislányként és kislányként meghatározott beszélő számára gyakran tegező formát javasol, ami nem helyes. A javasolt magázó formák egy része (pl.: *Üdvözlöm*) azonban a beszélő életkorának nem felel meg. A *Hello, asszonyom!* köszöntés pedig sem az udvariassági, sem pedig az életkori elvárásoknak nem felel meg.

Végül harmadik kvalitatív kiértékelésünk során 117, magyar kultúrával, történelemmel, nyelvvel kapcsolatos kérdésre adott válaszok tényyszerűségét hasonlítottuk össze. A tényyszerűséget először egy hármas skálán értékeltük:

- 0: a promptrá adott válasz teljesen helytelen
- 1: a promptrá adott válasz részben helyes, de tartalmaz ténybeli tévedéseket, vagy helyes, de nem a legrelevánsabb választ adta

- 2: a válasz tényyszerű és pontos, nem tartalmaz ténybeli tévedéseket

Fontos megjegyezni, hogy nem vettük figyelembe a válasz stilisztikai jellemzőit, a nyelvi helyességét, a formai megfelelőségét, csak azt vizsgáltuk, hogy a kérdésre helyesen válaszolnak-e a modellek. Az eredményeket az 5. táblázatban foglaltuk össze.

Témakör	Kérdések száma	PULI	LlumiX	Instruct	ParancsPuli	Sambalingo
Egyéb	19		61%		47%	66%
Földrajz	17		82%		85%	71%
Irodalom	26		67%		63%	65%
Művészet	14		54%		61%	43%
Nyelv	22		70%		48%	45%
Történelem	16		72%		41%	75%
	114		67%		57%	61%

5. táblázat. A modellek kiértékelésének eredménye a kulturális tárgyi tudást mérő adathalmazon.

A PULI LlumiX Instruct modell jó teljesítményt nyújtott a kulturális tárgyi tudást mérő kérdések megválaszolásában, és a legmagasabb átlagos pontosságot érte el (**67%**). Különösen jól szerepelt az *Irodalom* kategóriában, ahol a többi modellt meghaladva **67%-os** pontossággal válaszolt a kérdésekre, szemben a *ParancsPuli* 63%-os és a *Sambalingo* 65%-os eredményével. A *Nyelv* témakörben a PULI LlumiX Instruct modell 70%-os pontossága jóval felülmúlja a másik két modell alacsony, 48% és 45%-os pontosságát.

Fontos azonban megemlíteni, hogy bizonyos témakörökben más modellek is felülmúlták a PULI LlumiX Instruct modellt. A *Földrajz* kategóriában például a *ParancsPuli* érte el a legjobb eredményt **85%-kal**, amely kicsit meghaladja a PULI LlumiX Instruct modell 82%-os teljesítményét, míg a *Sambalingo* itt 71%-kal szerepelt. A *Történelem* kategóriában a *Sambalingo* bizonyult a legjobbnak, 75%-os pontossággal, ami kicsit felülmúlja a PULI LlumiX Instruct modell 72%-os eredményét. Az *Egyéb* kategóriában a *Sambalingo* vezet, **66%-kal**, míg a PULI LlumiX Instruct modell 61%-ot ért el ezen a területen.

6. Összegzés

Kutatásunk során kifejlesztettünk egy magyar nyelvre adaptált, Llama-2 alapú, folytatólagosan előtanított és utasításkövető finomhangolással továbbfejlesztett nagyméretű nyelvi modellt. A modell különböző hazai benchmarkokon, mint a HuCOLA, HuSST és HuRTE, kimagasló pontosságot ért el, felülmúlva a korábbi modelleket, és erősségeit kvalitatív elemzések is igazolták. Eredményeink rávilágítanak arra, hogy a transzferatanulás jelentősen növeli a modell nyelvi teljesítményét, lehetőséget nyújtva a magyar nyelvű nyelvtechnológiai feladatok pontosabb megoldására.

Hivatkozások

- Alves, D.M., Pombal, J., Guerreiro, N.M., Martins, P.H., Alves, J., Farajian, A., Peters, B., Rei, R., Fernandes, P., Agrawal, S., Colombo, P., de Souza, J.G.C., Martins, A.F.T.: Tower: An open multilingual large language model for translation-related tasks (2024)
- Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M.: The balanced accuracy and its posterior distribution. In: 2010 20th International Conference on Pattern Recognition. pp. 3121–3124 (2010)
- Chen, S., Wong, S., Chen, L., Tian, Y.: Extending context window of large language models via positional interpolation (2023), <https://arxiv.org/abs/2306.15595>
- Conover, M., Hayes, M., Mathur, A., Xie, J., Wan, J., Shah, S., Ghodsi, A., Wendell, P., Zaharia, M., Xin, R.: Free Dolly: Introducing the World’s First Truly Open Instruction-Tuned LLM (2023), <https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm>
- Csaki, Z., Li, B., Li, J., Xu, Q., Pawakapan, P., Zhang, L., Du, Y., Zhao, H., Hu, C., Thakker, U.: Sambalingo: Teaching large language models new languages (2024)
- Csaki, Z., Pawakapan, P., Thakker, U., Xu, Q.: Efficiently adapting pretrained language models to new languages (2023), <https://arxiv.org/abs/2311.05741>
- Dodé, R., Yang, Z.Gy.: Kulcsszógenerálás magyar nyelvű, hosszú szövegekből nagy nyelvi modellekkel. In: Berend, G., Gosztolya, G., Vincze, V. (szerk.) XX. Magyar Számítógépes Nyelvészeti Konferencia. pp. 257–267. Szegedi Tudományegyetem, online (2024)
- Ke, Z., Shao, Y., Lin, H., Konishi, T., Kim, G., Liu, B.: Continual pre-training of language models. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=m_GDIItaI3o
- Ke, Z., Shao, Y., Lin, H., Xu, H., Shu, L., Liu, B.: Adapting a language model while preserving its general knowledge. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (szerk.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 10177–10188. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022), <https://aclanthology.org/2022.emnlp-main.693>
- Kryscinski, W., Rajani, N., Agarwal, D., Xiong, C., Radev, D.: BOOKSUM: A collection of datasets for long-form narrative summarization. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (szerk.) Findings of the Association for Computational Linguistics: EMNLP 2022. pp. 6536–6558. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022), <https://aclanthology.org/2022.findings-emnlp.488>
- Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.W., Dai, A.M., Uszkoreit, J., Le, Q., Petrov, S.: Natural questions: A benchmark for question answering research. Transactions of the Association for Computational Linguistics 7, 452–466 (2019), <https://aclanthology.org/Q19-1026>

- Laki, L.J., Yang, Z.Gy.: Sentiment analysis with neural models for hungarian. *Acta Polytechnica Hungarica* 20(5), 109–128 (2023)
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive nlp tasks. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS '20, Curran Associates Inc., Red Hook, NY, USA (2020)
- Ligeti-Nagy, N., Ferenczi, G., Héja, E., Jelencsik-Mátyus, K., Laki, L.J., Vádász, N., Yang, Z.Gy., Váradi, T.: HuLU: magyar nyelvű benchmark adatbázis kiépítése a neurális nyelvmodellek kiértékelése céljából. In: *XVIII. Magyar Számítógépes Nyelvészeti Konferencia*. p. 431–446. JATEPress, Szeged (2022)
- Ligeti-Nagy, N., Héja, E., Laki, L.J., Takács, D., Yang, Z.Gy., Váradi, T.: Hát te mekkorát nőttél! - A HuLU első életévé új adatbázisokkal és webszolgáltatással. In: *XIX. Magyar Számítógépes Nyelvészeti Konferencia*. p. 217–230. JATEPress, Szeged (2023)
- Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Stanford Alpaca: An Instruction-following LLaMA model (2023), https://github.com/tatsu-lab/stanford_alpaca
- 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.: LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971* (2023a)
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardaş, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open Foundation and Fine-Tuned Chat Models (2023b)
- Xu, H., Liu, B., Shu, L., Yu, P.: BERT post-training for review reading comprehension and aspect-based sentiment analysis. In: Burstein, J., Doran, C., Solorio, T. (szerk.) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. pp. 2324–2335. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019), <https://aclanthology.org/N19-1242>
- Yang, Z.Gy., Dodé, R., Ferenczi, G., Hatvani, P., Héja, E., Madarász, G., Ligeti-Nagy, N., Sárosy, B., Szaniszló, Zs., Váradi, T., Verebélyi, T., Prószéky, G.: The first instruct-following large language models for hungarian. In: *2024 IEEE 3rd Conference on Information Technology and Data Science (CITDS) Proceedings*. pp. 247–252. University of Debrecen, Debrecen, Hungary (2024a)

- Yang, Z.Gy., Dodé, R., Ferenczi, G., Héja, E., Jelencsik-Mátyus, K., Kőrös, A., Laki, L.J., Ligeti-Nagy, N., Vadász, N., Váradi, T.: Jönnék a nagyok! BERT-Large, GPT-2 és GPT-3 nyelvmodellek magyar nyelvre. In: XIX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2023). pp. 247–262. Szegedi Tudományegyetem, Informatikai Intézet, Szeged, Hungary (2023a)
- Yang, Z.Gy., Laki, L.J., Váradi, T., Prószéky, G.: Mono- and multilingual gpt-3 models for hungarian. In: Text, Speech, and Dialogue. pp. 94–104. Lecture Notes in Computer Science, Springer Nature Switzerland, Plzeň, Czech Republic (2023b)
- Yang, Z.Gy., Ligeti-Nagy, N.: Improve performance of fine-tuning language models with prompting. Infocommunications Journal Special Issue on Applied Informatics, 62–68 (2023), <https://doi.org/10.36244/ICJ.2023.5.10>
- Yang, Z.Gy., Szilávik, S., Ligeti-Nagy, N.: Magyar nyelvű utasításkövető korpusz építése Stanford Alpaca promptok fordításával és lokalizálásával [Building a Hungarian instruct-following corpus by translating and localizing Stanford Alpaca prompts]. In: XX. Magyar Számítógépes Nyelvészeti Konferencia. pp. 243–255. Szegedi Tudományegyetem, Szeged (2024b)
- Zhu, W., Liu, H., Dong, Q., Xu, J., Huang, S., Kong, L., Chen, J., Li, L.: Multilingual machine translation with large language models: Empirical results and analysis (2023)