Középkori nyelvek feldolgozása ingyenes és kereskedelmi generatív nyelvmodellekkel

Pethő Gergely[1,2], Swaroop Krishna[3]

[1]Debreceni Egyetem, Egy Egészség Intézet
[2]Regensburgi Egyetem, Romanisztikai Intézet
[3]Debreceni Egyetem, Informatikai Kar

Kivonat A középkori szövegek számítógépes feldolgozása hagyományosan számos nehézségbe ütközik, amelyek főleg két okra vezethetők vissza. Az egyik a szükséges erőforrások hiánya akár szoftvertermékek (morfológiai és szintaktikai elemzők), akár adatok (elektronikus szótárak, kézzel annotált tanítóadatok) tekintetében. A másik a nyelvi sztenderd hiánya, ami kiterjed mind a szabályozott és következetes helyesírás hiányára, mind pedig „egy" adott nyelv számottevő szinkrón és diakrón változatosságára. Kutatásunkban azt vizsgáltuk, hogy a napjainkban divatos „kis" és „közepes" méretű ingyenes, saját hardveren futtatható generatív nyelvmodellek, amelyeket többnyelvű modern nyelvi adatokon tanítottak elő, mennyire alkalmasak történelmi nyelvek feldolgozására a nagy kereskedelmi modellekkel összevetve. Ehhez Albucasis sebészetről szóló tankönyvének arab eredetije, valamint annak latin, ófrancia és óokcitán nyelvű középkori változatai alapján e modellekkel készített gépi fordításokat értékeltünk ki a mű filológiai igényű angol fordítását referenciaként használva. Az eredmények segítenek annak megítélésében, hogy középkori nyelvek feldolgozásakor milyen generatív nyelvmodellek használatával érdemes próbálkozni a számítási erőforrásokat és a költségeket is figyelembe véve.
Kulcsszavak: generatív nyelvmodell, LLM, SLM, nyelvtörténet, alacsony erőforrású nyelvek

1. Bevezetés

A történelmi szövegek feldolgozásában a megszokott nyelvtechnológiai megoldások alkalmazása komoly nehézségekbe ütközik. A hagyományos szótár- és szabályalapú eszközök fejlesztésének és alkalmazásának fő akadályát a korai nyelvállapotok „szabályozatlansága", a nyelvi sztenderd hiánya jelenti. Ez kiterjed mind a helyesírási szabályok, illetve általánosan követett szokások hiányára, mind pedig arra, hogy a nyelvet egységesítő kulturális termékek (könyvnyomtatás, közoktatás, sajtó stb.) hiányában a szinkrón nyelvállapotokon belül jelentősen nagyobbak voltak a területi vagy akár egyéni eltérések, mint az újkortól kezdve. A szinkrón állapoton belüli változatosság mellett a diakrón eltérések is számottevőek, így még az egyazon nyelvtörténeti korszakban keletkezett írások nyelvi

jellemzői is drasztikus eltéréseket mutathatnak. Mindezen tényezők kizárják a hagyományos számítógépes nyelvfeldolgozási megoldások átfogó alkalmazását.

A gépi tanuláson, ezen belül különösen az újabb neurális nyelvmodelleken alapuló eszközök rendelkeznek olyan jellemzőkkel, amelyek alkalmassá teszik ezeket az ilyen értelemben „problémás" szövegek feldolgozására. Működésük nagyjából illeszkedő minták illesztésén nyugszik, így például egy generatív nyelvmodell olyan folytatást generál (azaz következő tokent, illetve iteratíven ilyenek sorozatát) egy adott prompthoz (azaz bemenetként kapott tokensorozathoz), amely hasonlít azokhoz a tokensorozatokhoz, amelyek a tanítókorpuszában a prompthoz hasonló, azonban vele rendszerint nem pontosan egyező tokensorozatok után következtek. Ennek köszönhetően a neurális nyelvmodellek jól ellenállnak a feldolgozandó szövegben megtalálható zajnak, írásmódbeli („helyesírási"), morfológiai, szóhasználatbeli anomáliáknak. Ez a képesség, amely a neurális nyelvmodelleket alkalmassá teszi a sztenderdtől eltérő nyelvváltozatokban íródott nyelvi termékek, így például a közösségi médiában funkcionálisan írástudatlan nyelvhasználók által előállított szövegek értelmezésére, segítségükre lehet a régi szövegek feldolgozásában is.

Ugyanakkor a korai történelmi nyelvállapotokkal kapcsolatban nehézséget jelent, hogy egyrészt kevés olyan nyelvmodell áll rendelkezésre, amely kifejezetten ilyen nyelvű szövegeken lett tanítva, másrészt a tanításra rendelkezésre álló szövegek mennyisége eleve erősen korlátozott még a gazdag nyelvemlékanyaggal rendelkező nyelvek esetében is. Az összes ómagyar szövegemlék (MGTSZ korpusz) terjedelme például mindössze 3 millió szövegszó, míg a középkori francia korpusz (BFM) mintegy 6 millió szövegszót foglal magában. Ez legalább 2-3 nagyságrenddel kevesebb egy használható neurális nyelvmodell tanításához szükséges adatmennyiségnél. A középkori nyelvek feldolgozására képes neurális nyelvmodellek fejlesztése így a jelenlegi architektúrák és tanítási elvek (azaz rekurrens vagy transzformerhálózatok gradiensmódszerrel, maszkolásos tanítással tanítva) mellett a közismert skálázási szabályokat figyelembe véve (Kaplan és mtsai, 2020; lásd még Hoffmann és mtsai, 2024) gyakorlatilag reménytelen.

Tanulmányunkban azt járjuk körül, hogy a döntően modern nyelvi szövegeken tanított nyelvmodellek alkalmasak-e középkori nyelvek feldolgozására. Más szóval azt vizsgáljuk, hogy történelmi szövegek mennyire sikeresen modellezhetőek eltorzított, zajos modern nyelvi szövegekként. E célkitűzéssel kapcsolatban felmerül ugyanakkor az a módszertani kérdés, hogy miként lehet számszerűen mérni és összehasonlítani az egyes modellek használhatóságát, tekintve, hogy nem állnak rendelkezésre (tudomásunk szerint) ilyen benchmarkok. Ezek hiányában a többnyelvű generatív nyelvmodellek egy jól ismert képességét aknáztuk ki nyelvtudásuk tesztelésére: az adott nyelvből angolra fordíttattunk velük. Bár vizsgálatunk eredményei közvetlenül csak a modellek fordítási képességéről adnak tájékoztatást, a generatív nyelvmodellek univerzális alkalmazhatóságával kapcsolatos tapasztalatok alapján abból indulhatunk ki, hogy arra is következtetni engednek, megfelelő promptolás vagy finomhangolás mellett adott modell más célokra, például morfológiai, szintaktikai címkézésre, bizonyos jelentéskategóriákba eső szavak (pl. tulajdonnevek) felismerésére stb. használhatóvá tehető-e.

Vizsgálatunk egy olyan, viszonylag nagy terjedelmű szövegen alapul, amelynek arab nyelvű eredetije a 10. század végén íródott, ennek a 13. században készült el latin nyelvű fordítása, majd feltehetőleg ez utóbbi fordítás alapján két vernakuláris fordítás ófrancia és óokcitán nyelvre. Ennek köszönhetően egyazon szöveg négy különböző középkori változata alapján készült gépi fordítások minőségét tudjuk összehasonlítani. Így lehetővé válik, hogy nemcsak a vizsgált nyelvmodellek fordítási képességét vessük össze, hanem azt is, hogy különböző középkori nyelvek egymáshoz képest mennyire sikeresen dolgozhatók fel velük.

2. Kapcsolódó irodalom

A nagy generatív nyelvmodelleknek a középkori, illetve általánosabban a történelmi nyelvek és nyelvállapotok feldolgozására való használata feltáratlan terület, nem tudunk átfogó kutatásról e kérdéskörrel kapcsolatban.

A klasszikus arab nyelv nagy generatív nyelvmodellekkel történő feldolgozására nincs példa a szakirodalomban. Az e nyelvvel foglalkozó legújabb szakirodalom csak kisebb, elsősorban modern arab köznyelvi anyagra előtanított encodernyelvmodellek alkalmazásáról számol be, amelyek tanítását kiegészítették korlátozott mennyiségű klasszikus arab szöveggel is (Malhas és Elsayed 2022, ElKoumy és Sarhan 2024), illetve találunk példát kifejezetten klasszikus arabra tanított BERT-modellre is (Inoue és mtsai, 2021).

Ókori nyelvek, többek között a latin gépi tanulási eszközökkel történő feldolgozásáról közölt a közelmúltban átfogó áttekintést Sommerschield és mtsai (2023). Generatív nyelvmodellek alkalmazására nem találunk példát. Volk és mtsai (2024) azt vizsgálja egy igen kis terjedelmű, mindössze 3000 szövegszós tesztanyagon, hogy kora újkori latin és német nyelvű írások fordítására mennyire alkalmas a GPT-4 a Google Fordítóval és más gépi fordító eszközökkel összevetve. Stüssi (2023) és Stüssi és Ströbel (2024) alapos, szisztematikus munka, amely különböző ChatGPT-verziók használhatóságát vizsgálja 16. századi latin szövegek tokenjeinek szófaji címkézésére.

Torres Aguilar (2022) túlnyomórészt modern anyagon tanított többnyelvű encodermodelleket használ középkori, többek között latin és ófrancia nyelvű szövegekben tulajdonnevek felismerésére.

Her és Kruschwitz (2024) és Ondrejová és Šuppa (2024) azt a kérdést vizsgálja, hogy modern nyelvek alacsony erőforrású nyelvjárásait (a német nyelv bajor, illetve a szlovák nyelv šariš dialektusát) mennyire sikeresen képesek feldolgozni generatív nyelvmodellek.

A jelen tanulmány közvetlen előzménye két konferencia-előadás, amelyek keretében ismertettük eljárásunkat és az ófranciára és óokcitánra vonatkozó, e tanulmányban is idézett eredményeinket (Pethő és mtsai, 2024b), illetve az arabra és latinra vonatkozókat (Pethő és mtsai, 2024a).

3. Anyagok

Vizsgálatunk alapjául a középkori Európában leginkább Albucasis néven ismert 10. századi andalúziai arab orvos 30 kötetes, *Kitáb at-taszríf* című orvosi enciklopédiájának utolsó, egyben legnagyobb terjedelmű, *A sebészetről és az eszközökről* című kötete szolgált. Albucasis *Sebészete* (ahogy a továbbiakban hivatkozunk rá) az egyik legfontosabb orvosi tankönyv volt a második évezred első évszázadaiban, jelentősége és hatása révén számos másolatban és több fordításban fennmaradt. Vizsgálatunk során azt elemeztük, hogy a *Sebészet* eredetijének és három különböző középkori európai fordításának különböző generatív nyelvmodellekkel készült angol nyelvű gépi fordításai mennyire jól közelítenek az arab mű kritikai kiadásában (Spink és Lewis, 1973) közölt, tudományos igénnyel készült angol nyelvű referenciafordításhoz. A *Sebészet* három könyvön belül 200 fejezetből áll. A terjedelem szövegszavakban mérve nyelvenként eltérő, `wc -w` szerint: arab 59873, latin 75582, francia 77120, okcitán 99528, angol 91283. A tokenszám egy találomra kiválasztott tokenizáló (a Google Gemma 2 9b-é) szerint: arab 126002, latin 145723, francia 122661, okcitán 154918, angol 110222.

Albucasis tankönyve klasszikus arab nyelven íródott. Ez a 9. század során nyelvtan- és szótárírók által kanonizált, onnantól kezdve az iszlám szent nyelveként változatlan formában fennmaradt nyelv mind szintaxisában, mind morfológiájában igen közel áll a 19. században kialakult modern arab köznyelvhez, amely összarab műveltségi nyelvként funkcionál napjainkig.

Az arab eredeti mellett a mű 1200 körül keletkezett latin, 13. századi ófrancia és 14. századi óokcitán nyelvű fordítását dolgoztuk fel. Albucasis más teljes középkori fordítása nem ismert. Ugyan a mű fordításainak szöveghagyományát még nem tárta fel a kutatás, a szakirodalom (pl. Trotter, 1999, Green, 2011) feltételezi, hogy mind az ófrancia, mind az óokcitán változat nem közvetlenül az arab eredetin, hanem a latin fordításon alapul.

A latin fordítás több mint 20 másolatban középkori kéziratként máig fennmaradt (ennek részleteiről lásd Green, 2011), kritikai kiadása nem készült. Latin elektronikus szövegként egy 1532-es nyomtatvány alapján a würzburgi egyetemen Irina Galynina által készített XML-dokumentumot használtunk. Az ófrancia fordítás kritikai kiadása Trotter (2005), ezen alapul a francia középkori korpusz (BFM) számára Alexei Lavrentiev által készített XML-változat, amelyet használtunk. Az óokcitán fordításnak több kritikai kiadása is készült, ezek közül a legpontosabb a szakma véleménye szerint Elsheikh (1992). Az ez alapján P.T. Ricketts által készített, majd Dominique Billy által TEI formátumra hozott XML-t használtuk a feldolgozás során. Az arab eredeti már említett kritikai kiadása alapján szintén Würzburgban Samer Alsaj és Azzam Hasan által készített XML-változattal dolgoztunk.

Amint már említettük, referenciafordításként az arab mű kritikai kiadásában (Spink és Lewis, 1973) szereplő angol fordítást használtuk. E változatokon kívül tudunk még az arab eredetinek egy 18. századi tudományos latin, egy 19. század végi modern francia, valamint egy, az 1980-as években készült filológiai orosz fordításáról, azonban ezeket nem használtuk.

4. Módszer

4.1. Előfeldolgozás

A négy középkori nyelv esetében az említett XML-dokumentumokból indultunk ki. A referenciafordításként használt Spink és Lewis-féle angol nyelvű fordításnak nem rendelkeztünk digitalizált változatával, ezért ezt magunk készítettük el: A kritikai kiadás PDF-változatából a PyMuPDF (`https://github.com/pymupdf/PyMuPDF`) szoftver segítségével kinyertük az angol nyelvű oldalak meglévő OCR-szövegét, automatikusan eltávolítottuk a lábjegyzeteket, oldalszámokat, végül részben automatizáltan, részben kézzel olyan XML-szerkezetet alakítottunk ki a szövegben, amely pontosan tükrözte az arab változatnak megfelelő fejezetegységeket. Ugyanezen fejezetek az okcitán kéziratban (és az XML-ben) címmel és fejezetszámmal jelölve voltak, azonban a dokumentum nem volt ezek mentén XML-elemekre tagolva, így ezeket hozzá kellett adni. A francia XML ugyan már tartalmazott fejezetelemeket, ezek számozása és határai azonban eltértek a másik három nyelvi változat egységes szerkezetétől, ezért ezeket kézzel hozzáigazítottuk az utóbbiakhoz. Végeredményként azonos szerkezetű, azonos módon tagolt, fejezetszinten illesztett XML-eket kaptunk.

Az előfeldolgozás következő lépése a mondatokra tagolás volt. A meglévő XML-dokumentumok mondatelemeket nem tartalmaztak. A bekezdéselemeket a négy indoeurópai nyelv esetében az XML-ben szereplő központozás alapján a SpaCy angol, illetve (a három másik nyelvben) többnyelvű modelljével bontottuk automatikusan mondatokra. Az ófrancia kritikai kiadásban annyira hosszú, esetenként oldalnyi egységek voltak mondatvégi ponttal jelölve, hogy itt a SpaCy tagolása nem volt elegendő, így az utóbbival kapott „mondatokat” tovább bontottuk a kettőspontok és pontosvesszők mentén. Az arab kritikai kiadás és XML nem tartalmazott mondatvégi központozást, így ott ugyanez nem volt járható. Különböző megoldásokkal próbálkoztunk, többek között azzal, hogy LLM-ekkel illesszük a Spink és Lewis-féle, már mondatokra tagolt fordításhoz az arab szöveget, illetve tagoltassuk mondategységekre a bekezdéseket, de végül kézi feldolgozás mellett döntöttünk. Az XML-t arab anyanyelvű munkatársunk, Esra' Abdelzaher bontotta a klasszikus arab hagyomány értelmében mondatnak minősülő egységekre. Amint utólag kiderült, ez olyan sajátos egységekhez vezetett, amelyekben a mondat- és tagmondathatárok sok esetben markánsan eltértek az indoeurópai nyelvek megfelelő mondatainak tagolásától, így azokra sokszor nem illeszkedtek pontosan.

A mondattagolás nyomán az angolban 3486, az arabban 2343, a latinban 4055, a franciában 3621, az okcitánban 3097 mondategységet kaptunk.

4.2. Mondatszintű illesztés

A négy történelmi szöveget mondatszinten automatikusan az angol referenciafordításhoz illesztettük fejezetenként haladva. Ehhez a SentAlign alkalmazást használtuk (Steingrímsson és mtsai, 2023). Mivel nem lehettünk biztosak abban, hogy a SentAlign által használt többnyelvű mondatbeágyazás-modell elég jól tud

a négy történelmi nyelven ahhoz, hogy sikeresen illeszteni tudja a szövegeket, előbb a történelmi szövegeket egy tetszőleges nagy kereskedelmi nyelvmodellel (történetesen kényelmi okokból arab: Gemini Pro; latin, okcitán: GPT-4o; francia: Claude Sonnet) szigorúan mondategységenként lefordíttattuk angolra, majd az így kapott, előzetes ellenőrzésünk alapján általunk jó minőségűnek ítélt gépi fordítást illesztettük a szintén angol referenciához. A SentAlign futása mintegy 3 percet vett igénybe nyelvpáronként. Ugyan kiváló minőségű illesztéseket hozott létre, mindemellett minden nyelvpár esetében illesztetlenül hagyott olyan mondatokat, amelyek ténylegesen valamelyik oldalon 3-nál vagy 4-nél több mondategységet tartalmazó 1:n vagy m:n biszegmenst alkottak. Ezért minden nyelvpár esetében az illesztéshez használt angol gépi fordítást és a referenciafordítást összevetve ellenőriztük az összes illesztetlen mondategységet, és szükség esetén kézzel javítottuk az illesztést.

Az illesztés nyomán az arab-angol párra 2173, a latinra 2881, a franciára 2517, az okcitánra 2806 biszegmenst kaptunk.

4.3. A gépi fordításhoz használt modellek

A 1. táblázatban felsorolt modelleket értékeltük ki.

Modell neve	**Paraméterek száma (mrd.)**	**Ingyenes?**
Aya 23	8 és 35	+
Aya Expanse	8	+
Claude 3.5 Sonnet	ismeretlen	–
Command R	35	+
Gemini 1.5 Flash 8b, Flash, Pro	8, illetve ismeretlen	–
Gemma	9	+
Gemma 2	9 és 27	+
Granite 3 Dense	2 és 8	+
GPT-4o (ChatGPT)	ismeretlen	–
Llama 3	8 és 70	+
Llama 3.1	8	+
Llama 3.2	3	+
Mistral	7	+
Mistral-Nemo	12	+
Mistral-Small	22	+
Mixtral	8 x 7	+
Nemotron-Mini	4	+
Phi-3	3,8 és 14	+
Phi-3.5	3,8	+
Qwen2	7	+
Qwen2.5	7, 14 és 32	+

1. táblázat. Kiértékelt modellek

Mindhárom fontos kereskedelmi modell nagy változata (Claude 3.5 Sonnet, Gemini 1.5 Pro, ChatGPT GPT-4o) mellett számos különböző méretű modellt vizsgáltunk, amelyek az Ollama alkalmazással (https://ollama.com/) könnyedén telepíthetőek és futtathatóak GPU-val (grafikus processzorral) felszerelt személyi számítógépeken. Az általunk tesztelt modellek mai mércével mérve többnyire a kis-közepes méretkategóriába tartoznak, a legkisebbek (2-4 milliárd paraméter) pedig kifejezetten ún. kis nyelvmodellnek (SLM) minősülnek.

A ChatGPT és a Claude esetében a forrásnyelvi szövegrészeket az ingyenes, böngészőben elérhető felületen adtuk át a modelleknek, azokat kézzel a csevegőmezőbe másolva, majd a generált fordítást a böngészőből egy szöveges dokumentumba másoltuk. A Geminit a Google API-n keresztül, ingyenes Google Cloud kreditekkel fizetve használtunk. Minden más modellt az Ollama alkalmazással futtattunk részben lokális asztali számítógépeken, részben a HUN-REN felhőben rendelkezésünkre álló két V100 GPU-s virtuális gépen.

A 70 milliárd paraméteres Llama 3 kivételével minden ingyenes modellt az Ollama modellkönyvtárban alapértelmezett 4 bites kvantált változatban (`Q4_0`) használtuk, ezen belül az utasításkövetésre hangolt (instruct) változatban. Néhány találomra kiválasztott modell esetében kiértékeltük próbaképpen a 8 bites kvantált és a teljes 16 bites lebegőpontos (`fp16`) változatot is, de miután semmilyen eltérést nem tapasztaltunk a fordítások minőségében a 4 bites kvantált modellekhez képest, az utóbbiakkal vittük végig a vizsgálatot.

A kiértékelt modellek kiválasztása tekintetében döntő szempontnak tekintettük, hogy ezek futtatásához ne legyen szükség szuperszámítógépes környezetre, hanem egy viszonylag olcsó, néhány száz eurós kereskedelmi grafikus kártyán is működjenek. Ezek körében tipikus RAM-felszereltségek a következő modellek futatására elegendőek: 8 GB memóriájú grafikus kártyákon a 8 milliárdig terjedő paraméterszámú modellek, 12 GB GPU-memóriával a 14 milliárdig terjedőek, 24 GB-tal (két darab 12 GB-os grafikus kártyával felszerelt PC-n) a 32 milliárdig terjedőek futtathatóak. A 70 milliárd paraméteres Llama 3 modell csak legfeljebb 2 bites kvantált változatban fut az általunk vizsgált legnagyobb grafikus memóriával (32 GB) rendelkező konfiguráción, így azt teszteltük. (Számunkra tökéletesen lényegtelen, hogy e modell 4 bites kvantált változata milyen teljesítményt nyújtott volna, ugyanis annak használatához bő 40 GB GPU-memóriára lenne szükség, ami lokális hardveren gyakorlatilag elérhetetlen.) Az ennél nagyobb modellek, pl. a 72b-s Qwen2 legkisebb kvantált változata sem futott 32 GB-on, így a 70b-nél nagyobb modelleket egyáltalán nem vizsgáltuk. Ugyan az Ollama úgy is képes futtatni nyelvmodelleket, hogy a modell nem fér el teljes mértékben a GPU-RAM-ban, és ilyen esetekben részben CPU-n, részben GPU-n futtatja a modellt, annak áteresztőképessége olyan drasztikusan csökken, hogy praktikusan nem használhatóak. Albucasis *Sebészet*-ének egy teljes fordítása 100%-ban GPU-n modelltől és hardvertől függően tipikusan 1-4 órát vesz igénybe.

A modellek kiértékelésére általánosan használt sztenderd szemantikai, matematikai stb. benchmarkokon (pl. *MMLU,* Hendrycks és mtsai, 2021a; *GPQA,* Rein és mtsai, 2023; *HellaSwag,* Zellers és mtsai, 2019; *Winogrande,* Sakaguchi és mtsai, 2021; *MATH,* Hendrycks és mtsai, 2021b stb. stb.) az az általános ta-

pasztalat és gyakorlat (vö. pl. Zhao és mtsai, 2024, Yang és mtsai, 2024), hogy a hasonló méretű modellek egymással tudnak versenyezni, és a jóval nagyobb modellek (pl. 14 milliárd szemben a 8 paraméterrel) mért teljesítménye világosan felülmúlja a kisebbekét. Előzetesen azt vártuk, hogy ez a középkori nyelvek feldolgozására is érvényes lesz.

Az ingyenesen használható modellek egy részét kifejezetten többnyelvűként hirdetik a készítőik, ilyen az Aya (23 nyelv), a Qwen 2 (29), a Qwen 2.5 (konkrét szám nélkül, többek között arab és francia), a Llama 3 család (30), a Mixtral (5, többek között francia, olasz, spanyol), a Command R („több mint 10"). A többi modell feltehetőleg erőteljesen angol központú tanítóanyagon alapul. Feltételeztük, hogy a dedikáltan többnyelvű modellek jelentősen jobban teljesítenek a többihez képest.

Amint látni fogjuk, az említett előzetes várakozások egyikét sem igazolták végül a mérések.

4.4. Promptolás

Mind a négy nyelvhez egy-egy angol nyelvű kezdő promptban összefoglaltuk a feladatot: Albucasis középkori sebészeti tankönyvének, illetve latin stb. nyelvű fordításának mondatait kell lefordítani angolra. A modellnek a megadott szöveg fordításával kell válaszolnia, semmi mást ne fűzzön hozzá. Az ófrancia és óokcitán fordítás kezdő promptja ezenkívül egy 4-shot példát is tartalmazott, tehát négy-négy ófrancia, illetve óokcitán mondatot, valamint mindegyik mondat után annak egy tetszőleges kereskedelmi LLM-mel (történetesen a GPT-4o-val) készített jó minőségű angol fordítását. A kezdő promptot követően a fordítandó mondategységeket egyenként adtuk át az Ollama Python-könyvtárával a nyelvmodellnek, válaszként egysoros fordítást vártunk. Ha nem egyetlen sorral válaszolt, a kontextust újra átadtuk legfeljebb három alkalommal, majd ha még ezután sem kaptunk egysoros választ, a kapott válasz első sorát tekintettük a modell által adott fordításnak. Ilyen esetekben, valamint akkor, ha a csevegés előzményei (tehát a felhasználó és a modell „hozzászólásai" páronként egymás után) hosszúsága meghaladta az 1000 tokent az adott nyelvmodell tokenizálója szerint, lenulláztuk a kontextust, és újból a kezdő prompttal, majd a soron következő fordítandó mondattal indítottuk a beszélgetést.

A Gemini modellek promptolása ettől az Ollamáétól eltérő API miatt különbözött, ott az instrukciókat rendszerpromptként adtuk át, kontextusként ezt követte a megelőző öt felhasználó–modell hozzájáruláspár, majd ezt zárta a következő fordítandó mondat.

A ChatGPT-nek és a Claude-nak egy a fentiekhez hasonló bevezető instrukció után kb. 3000-4000 karakternyi részt adtunk át egyszerre, ami tipikusan egy-egy fejezetnek felelt meg. Soronkénti fordítást kértünk, utólag a kimenet sorainak számát ellenőriztük, és ha nem egyezett a bemenetével, újra átadtuk ugyanazokat a sorokat fordításra. A ChatGPT esetében két, a Claude esetében minden egyes ilyen szakasz fordítása után új chatet kezdtünk. A ChatGPT és Claude esetében eltekintettünk az API használatától, mert ez pénzbe került volna, és a Google-lel ellentétben ezek a cégek nem adtak használható mennyiségű ingyen kreditet.

4.5. Kiértékelés

A három nagy kereskedelmi nyelvmodellel minden középkori szöveg (arab, latin, ófrancia, óokcitán) minden mondatához egy-egy angol fordítást generáltunk, majd az illesztett biszegmensek minőségét összesítve értékeltük ki.

Az óokcitánhoz minden egyéb modellel 10-10 gépi fordítást generáltunk a teljes szövegre, majd minden ilyen fordítást külön-külön kiértékeltük, hogy megállapítsuk, mekkora szórás tapasztalható az eredményekben. Azt találtuk, hogy modellenként a 10 fordítás között nem volt tapasztalható érdemi eltérés, pl. a BLEU-értékek között mindössze 0,2-0,3 pontnyi különbség jelentkezett, míg a különböző modellek fordításai között szignifikáns és nagy mértékű eltéréseket találtunk. Következésképpen nem lett volna haszna a további nyelvekhez is 10-10 fordítást generálnunk, ezért azokból csak 3-3 teljes fordítás készült.

A modellek által előállított kimenetek utófeldolgozására szándékosan nem tettünk kísérletet sem. Megfigyeltünk ugyan például olyan anomáliákat, hogy egyes modellek kéretlenül emojikat, díszítő elemeket, szögletes zárójelbe tett kommentárokat és hasonlókat tettek a válaszukba, és ezek egy részét könnyedén eltávolíthattuk volna. Ettől eltekintettünk annak érdekében, hogy az értékelés során kapott pontszámok azt is tükrözzék, ha egy modell rendszeresen oda nem tartozó szemetet tett a fordításba. A kapott pontszámok így nemcsak a modell fordítási képességét, hanem egyszersmind utasításkövetését is jellemzik.

A gépi fordítások és a referenciafordítás mondategységei alkotta biszegmenseket számos különböző fordításminőség-metrika alapján értékeltük. A SacreBLEU (Post, 2018) szoftverrel BLEU pontszámot, emellett egyéb szoftverekkel a fordítások NIST (Doddington, 2002), METEOR (Banerjee és Lavie, 2005), ROUGE (Lin, 2004), valamint BLEURT (Sellam és mtsai, 2020, Pu és mtsai, 2021) és BERTScore (Zhang és mtsai, 2020) pontszámát számoltattuk ki. A két utóbbi BERT (illetve tágabban encodermodell) alapú, a szemantikai egyezést számszerűsíteni próbáló mérőszám, míg a többi szóunigramoknak és -bigramoknak a gépi fordítás (hipotézis) és a referencia közötti egyezését, valamint relatív elhelyezkedését méri. A nagyszámú metrika mellett azért döntöttünk, mert látni akartuk, hogy azonos módon rangsorolják-e a különböző modellek kimenetét.

A vizsgálatunkhoz felhasznált valamennyi anyagot és szoftvert nyilvános repozitóriumban tettük elérhetővé: `https://gitlab.hadw-bw.de/gpetho/albuc_translate`. Ez az anyag a jövőben új modellek kiadása után könnyedén felhasználható lesz azoknak az általunk vizsgált négy középkori nyelven történő benchmarkolására.

5. Eredmények

A kiértékelések eredményét az 1. ábra összegzi. Minden kék vonal átlagosan 20-30 óra, minden más színű vonal kb. 5-10 óra GPU-idő árán előállított gépi fordítást jellemez. A három általunk kiszámított ROUGE pontszám közül csak a bigramok egyezését mérő ROUGE-2-ét közöljük, mivel a ROUGE-1 és a ROUGE-L diagramjai egymással majdnem pontosan egyeztek, valamint a modellek közötti

különbségek ezeken kevésbé egyértelműen rajzolódtak ki. A különböző metrikákkal kapott számokból egyező tendenciák olvashatóak le.

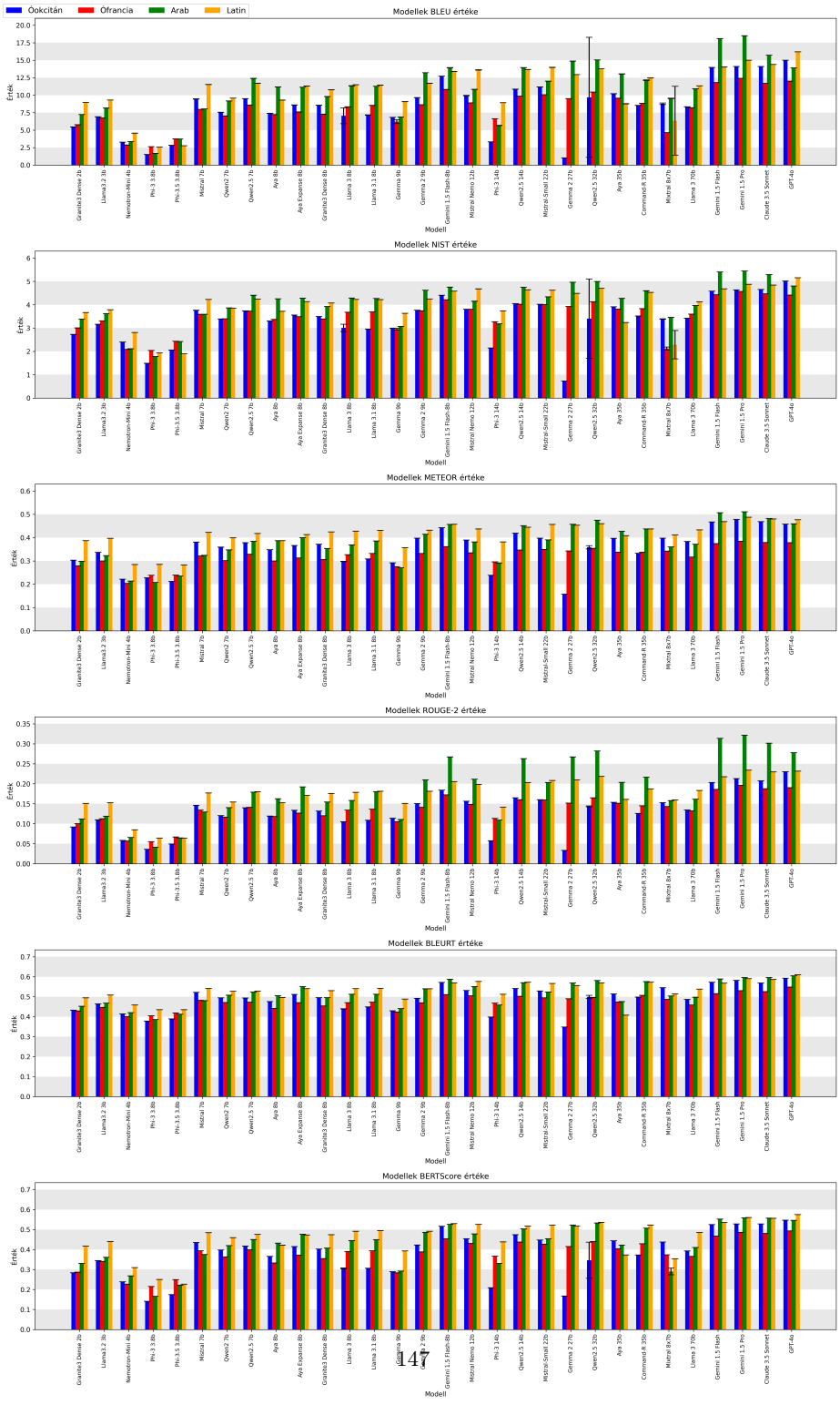
A 4 milliárd paraméterig terjedő tipikus SLM-ek közül a Granite3 és a Llama 3.2 látványosan felülmúlja a náluk jóval nagyobb Nemotron-Mini és Phi-3 modelleket, de ezzel együtt is gyengék a kis-közepes modellekhez mérten.

A 7-9 milliárd paraméter közötti modellek teszik ki a teljes mezőny kb. felét. Itt négy olyan párt is találunk, amelyeket egy modell és frissített változata alkotnak. Míg a Llama 3 és a Llama 3.1, illetve az Aya és az Aya Expanse teljesítménye között minimális eltérést tapasztalunk, a Gemma 2 a Gemmával, illetve a Qwen2.5 a Qwen2-vel szemben drasztikusan jobb lett mind a négy nyelvben, pedig ezek kiadása között mindössze 3-4 hónap telt el. Feltűnő a Gemini 1.5 Flash-8b teljesítménye, amely ugyan a Google által a Gemma modellek elnevezésében követett gyakorlattal összhangban valójában szinte biztosan nem 8, hanem 9 milliárd paramétert tartalmaz, de ezzel együtt is nemcsak a saját méretkategóriájából emelkedik ki látványosan, hanem a kétszer, háromszor ekkora ingyenes modelleket is felülmúlja a nyelvek többségében.

A 10 milliárd paraméter feletti kategóriában a legfeltűnőbb a hatalmas Llama 3 70b-nek még a tizedakkora modellekkel képest is igen gyenge szereplése. Az egyetlen vizsgált mixture-of-experts modell, a Mixtral 8x7b teljesítménye is kiábrándító. A nagy Phi-3 modell – SLM testvéreihez hasonlóan – gyakorlatilag használhatatlan, a kis-közepes méretű modellekhez képest is jóval gyengébb teljesítményt nyújtott. A 12 GB-os GPU-n kényelmesen futó, hasonló méretű Mistral-Nemo és Qwen2.5 14b nagyjából fej fej mellett, viszonylag jól teljesít, bár utóbbi következetesen minimálisan jobb az előbbinél, kivéve az arab feldolgozásában, ahol jelentősen jobb. A 20 és 40 milliárd paraméter közötti modellek általában minimálisan jobb eredményeket adtak az előbbiekhez képest.

A három nagy kereskedelmi nyelvmodell összességében hasonló teljesítményt nyújtott. Az arab és az ófrancia fordításában a Gemini, a latinban és az óokcitánban a GPT-4o teljesített legjobban. A két magas erőforrású nyelv tekintetében a legjobb ingyenes modellek pontszámai megközelítik, esetenként (vö. GPT-4o arab) felül is múlják a kereskedelmi LLM-ek eredményeit. Ugyanakkor az adott nyelvben legsikeresebb LLM látványosan pontosabban fordított, mint a legsikeresebb ingyenes modell (arab: Qwen2.5 32b és Gemini 1.5; latin: Mistral-Small és GPT-4o). Az alacsony erőforrású nyelvek tekintetében minden kereskedelmi modell messze felülmúlta a legsikeresebb ingyenes modellt (Qwen2.5 32b).

Fontos hangsúlyozni, hogy a referenciafordítás az arab eredeti alapján készült, így törvényszerűen azt tükrözi a legpontosabban. A latin fordítás viszonylag pontos, bár a fordító egyes arab kifejezéseket, amelyeknek nem volt kézenfekvő latin megfelelője, esetleg amelyeket nem értett, meghagyott latin betűkre átírt arab szavakként (vö. Trotter, 1999, 360–361), valamint azonosíthatóak kihagyások és félreértések is (i.m. 364). Az óokcitán és az ófrancia ezeket a rontásokat átvette, azonban ettől eltekintve az óokcitán szöveg jóval pontosabban megfeleltethető az arab eredetinek, mint az ófrancia. Az utóbbi fordítója feltűnően sok helyen pontatlanul dolgozott, sok részt kihagyott, leegyszerűsített. Mindebből következik, hogy ideális fordításminőség mellett az arab alapján készült fordításnak kell



1. ábra: Gépi fordítások kiértékelésének eredményei

legközelebb állnia a referenciához, ezt kell követnie a latinnak, az óokcitánnak, majd jelentősen lemaradva az ófranciának. Az elméletileg lehetséges optimális fordítások közötti viszonyt legpontosabban a Gemini BLEU eredményei tükrözik, illetve a sikeresebb modellek ROUGE értékei.

Az eredmények ugyanakkor semmit nem árulnak el arról, hogy a gépi fordítások abszolút értelemben mennyire jók, így például egy 15-ös BLEU-t mutató latinból készült fordítás mennyire pontos és érthető. Ugyan szisztematikus kvalitatív elemzést nem végeztünk, szúrópróbaszerűen nyelvenként 100-200 biszegmenst ellenőriztünk, és azt találtuk, hogy a kereskedelmi LLM-ek fordításminősége minden középkori nyelvre jól használható, nem marad el jelentősen egy emberi fordítótól elvárható minőségtől. A 10 alatti BLEU-t mutató fordítások ellenben összességében használhatatlanok voltak.

6. Összegzés

Eredményeink azt mutatják, hogy középkori latin és arab nyelvű szövegek számítógépes nyelvészeti feldolgozásában jól használhatóak a megfizethető GPU-val felszerelt asztali számítógépen is futtatható közepes méretű nyelvmodellek, így ezek feltehetőleg jól finomhangolhatóak lennének például címkézési feladatokra. Ugyanakkor tekintettel arra, hogy a Google által adott 300 dollár ingyen kreditből a történelmi korpuszokhoz képest igen nagy mennyiségű kimenet generálható a Gemini 1.5 Flash modellel, ennek használata észszerű alternatívája lehet az ingyenes modelleknek nyelvtörténeti NLP-alkalmazások fejlesztésekor.

Az alacsony erőforrású középkori nyelvek feldolgozása terén ugyanakkor az ingyenes modellek nem tudnak versenyezni a kereskedelmi LLM-ekkel. A 10 milliárd paraméternél kisebb modellek egyértelműen nem alkalmasak ezek értelmezésére. Az ugyanakkor nem derül ki egyértelműen az értékekből, hogy az ezeknél világosan jobban fordító közepes méretű Qwen2.5 és Mistral modellek megfelelő finomhangolással alkalmassá tehetők-e például óokcitán szövegek jó minőségű szófaji címkézésére, amire előzetes kísérleteink alapján a GPT-4o zero-shot alapon, finomhangolás nélkül is képes.

Köszönetnyilvánítás

Köszönjük Esra' Abdelzaher intenzív közreműködését az arab szövegváltozat előfeldolgozásában, ezen belül különösen az arab XML általunk használt módosított verziójának kialakításában. Christina Hodeib szintén hasznos megjegyzésekkel segítette az arab változat feldolgozását.

Hálásak vagyunk a Nyelvtudományi Kutatóközpont Lexikai Tudásreprezentáció Kutatócsoportjának, hogy díjmentesen engedélyezte számunkra a rendelkezésére álló GPU-s számítási kapacitás használatát e kutatás keretében.

A *Lexikai tudásreprezentáció* és a *Harmadik generációs szekvenálási adatok bioinformatikai elemzése* projektek nevében köszönetet mondunk a HUN-REN Cloud (lásd: Héder és mtsai, 2022; https://science-cloud.hu/) használatáért, ami hozzájárult a publikált eredmények eléréséhez.

Hivatkozások

Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. pp. 65–72. ACL, Ann Arbor, Michigan (Jun 2005), https://www.aclweb.org/anthology/W05-0909

Doddington, G.: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proceedings of HLT '02. p. 138–145. Morgan Kaufmann, San Francisco (2002)

ElKoumy, M., Sarhan, A.: Improving ranking-based question answering with weak supervision for low-resource Qur'anic texts. Artificial Intelligence Review 58(1), 17 (2024)

Elsheikh, M.S. (szerk.): Abū'l Qāsim Halaf Ibn Abbās az-Zahrāwī, La Chirurgia. Versione occitanica della prima metà del Trecento. nill, Firenze (1992)

Green, M.H.: Moving from philology to social history: The circulation and uses of Albucasis's Latin surgery in the Middle Ages. In: Glaze, F.E., Nance, B. (szerk.) Between text and patient: the medical enterprise in Medieval & Early Modern Europe, pp. 331–372. SISMEL, Edizioni del Galluzzo, Firenze (2011)

Héder, M., Rigó, E., Medgyesi, D., Lovas, R., Tenczer, S., Török, F., Farkas, A., Emődi, M., Kadlecsik, J., Mező, G., Pintér, Á., Kacsuk, P.: The past, present and future of the ELKH cloud. Információs Társadalom 22(2), 128 (aug 2022), https://doi.org/10.22503%2Finftars.xxii.2022.2.8

Hendrycks, D., és mtsai: Measuring massive multitask language understanding (2021a), https://arxiv.org/abs/2009.03300

Hendrycks, D., és mtsai: Measuring mathematical problem solving with the MATH dataset (2021b), https://arxiv.org/abs/2103.03874

Her, W., Kruschwitz, U.: Investigating neural machine translation for low-resource languages: Using Bavarian as a case study. In: Melero, M., Sakti, S., Soria, C. (szerk.) Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024. pp. 155–167. ELRA and ICCL, Torino, Italia (May 2024), https://aclanthology.org/2024.sigul-1.20

Hoffmann, J., és mtsai: Training compute-optimal large language models. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. pp. 30016–30030. NIPS '22, Curran Associates, Red Hook (2024), https://dl.acm.org/doi/10.5555/3600270.3602446

Inoue, G., és mtsai: The interplay of variant, size, and task type in Arabic pre-trained language models. In: Habash, N., és mtsai (szerk.) Proceedings of the Sixth Arabic Natural Language Processing Workshop. pp. 92–104. Association for Computational Linguistics, Kyiv (Apr 2021), https://aclanthology.org/2021.wanlp-1.10

Kaplan, J., és mtsai: Scaling laws for neural language models. CoRR abs/2001.08361 (2020), https://arxiv.org/abs/2001.08361

Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Lin-

guistics, Barcelona, Spain (Jul 2004), https://www.aclweb.org/anthology/W04-1013

Malhas, R., Elsayed, T.: Arabic machine reading comprehension on the holy Qur'an using CL-AraBERT. Information Processing & Management 59(6), 103068 (2022)

Ondrejová, V., Šuppa, M.: Can LLMs handle low-resource dialects? A case study on translation and common sense reasoning in Šariš. In: Scherrer, Y., Jauhiainen, T., Ljubešić, N., és mtsai (szerk.) Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024). pp. 130–139. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024), https://aclanthology.org/2024.vardial-1.11

Pethő, G., Abdelzaher, E., Krishna, S.: Evaluating generative language models on historical texts: A comparative study of Arabic, Latin and two low-resource medieval languages. In: Workshop on Natural language and AI: New perspectives for linguistic studies at ÖLT 2024, Innsbruck. Innsbruck (Dec 2024a)

Pethő, G., Menon, R., Schöffel, M., Tittel, S.: Comparing the ability of generative language models to understand medieval languages. In: 2nd International Conference on Data & Digital Humanities: Generative Artificial Intelligence for Text and Multimodal Data. Braga (Dec 2024b)

Post, M.: A call for clarity in reporting BLEU scores. In: Proceedings of the Third Conference on Machine Translation: Research Papers. pp. 186–191. Association for Computational Linguistics, Belgium, Brussels (Oct 2018), https://www.aclweb.org/anthology/W18-6319

Pu, A., Chung, H.W., Parikh, A.P., Gehrmann, S., Sellam, T.: Learning compact metrics for MT. In: Proceedings of EMNLP (2021)

Rein, D., és mtsai: GPQA: A graduate-level Google-proof Q&A benchmark (2023), https://arxiv.org/abs/2311.12022

Sakaguchi, K., Bras, R.L., Bhagavatula, C., Choi, Y.: WinoGrande: an adversarial winograd schema challenge at scale. Commun. ACM 64(9), 99–106 (Aug 2021), https://doi.org/10.1145/3474381

Sellam, T., Das, D., Parikh, A.P.: BLEURT: Learning robust metrics for text generation. In: Proceedings of ACL (2020)

Sommerschield, T., Assael, Y., Pavlopoulos, J., és mtsai: Machine learning for ancient languages: A survey. Computational Linguistics pp. 703–747 (Sep 2023), https://aclanthology.org/2023.cl-3.5

Spink, M.S., Lewis, G. (szerk.): Albucasis on Surgery and Instruments: A Definitive Edition of the Arabic Text with English Translation and Commentary. Publications of the Wellcome Institute of the History of Medicine; 12, The Wellcome Institute of the History of Medicine, London (1973)

Steingrímsson, S., Loftsson, H., Way, A.: SentAlign: Accurate and Scalable Sentence Alignment. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. ACL, Singapore (Dec 2023)

Stüssi, E.: Part-of-Speech Tagging on 16th-Century Latin. Institut fur Computerlinguistik, Universität Zürich (2023)

Stüssi, E., Ströbel, P.: Part-of-speech tagging of 16th-century Latin with GPT. In: Bizzoni, Y., Degaetano-Ortlieb, S., Kazantseva, A., Szpakowicz, S. (szerk.) Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024). pp. 196–206. ACL, St. Julians, Malta (Mar 2024), https://aclanthology.org/2024.latechclfl-1.18

Torres Aguilar, S.: Multilingual named entity recognition for medieval charters using stacked embeddings and BERT-based models. In: Sprugnoli, R., Passarotti, M. (szerk.) Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages. pp. 119–128. European Language Resources Association, Marseille, France (Jun 2022), https://aclanthology.org/2022.lt4hala-1.17

Trotter, D.A.: Arabic surgery in Eastern France and in the Midi: The Old French and Occitan versions of the Chirurige d' Albucasis. Forum for Modern Language Studies XXXV(4), 358–371 (1999)

Trotter, D. (szerk.): Albucasis, Traitier de Cyrurgie: Edition de la traduction en ancien français de la Chirurgie d'Abu'l Qasim Halaf Ibn Abbas al-Zahrawi du manuscrit BNF, français 1318. Niemeyer (2005)

Volk, M., Fischer, D.P., Fischer, L., és mtsai: LLM-based machine translation and summarization for Latin. In: Sprugnoli, R., Passarotti, M. (szerk.) Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024. pp. 122–128. ELRA and ICCL, Torino, Italia (May 2024), https://aclanthology.org/2024.lt4hala-1.15

Yang, A., és mtsai: Qwen2 technical report (2024), https://arxiv.org/abs/2407.10671

Zellers, R., és mtsai: HellaSwag: Can a machine really finish your sentence? (2019), https://arxiv.org/abs/1905.07830

Zhang, T., Kishore, V., Wu, F., és mtsai: BERTScore: Evaluating text generation with BERT. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=SkeHuCVFDr

Zhao, J., és mtsai: LoRA Land: 310 fine-tuned LLMs that rival GPT-4, A technical report (2024), https://arxiv.org/abs/2405.00732