Ekaia, ale berezia 2025, 13-27
https://doi.org/10.1387/ekaia.26338

ekaia
ZIENTZIA eta TEKNOLOGIA ALDIZKARIA

ISSN 0214-9001 – eISSN 2444-3255

Latxa euskarazko hizkuntza-eredua

(Latxa Language Model for Basque)

Naiara Pérez, Julen Etxaniz, Óscar Sainz, Itziar Aldabe, German Rigau, Eneko Agirre, Ahmed Salem, Aitor Ormazabal, Mikel Artetxe, Aitor Soroa*

HiTZ Hizkuntza Teknologiako Euskal Zentroa-Ixa,
Euskal Herriko Unibertsitatea (UPV/EHU)

LABURPENA: Artikulu honetan Latxa hizkuntza-ereduak (HE) aurkeztuko ditugu, egun euskararako garatu diren HE handienak. Latxa HEek 7.000 miloi parametrotik 70.000 milioira bitartean dituzte, eta ingeleseko LLama 2 ereduetatik eratorriak dira. Horretarako, «aurreikasketa jarraitua» izeneko prozesua gauzatu da LLa- ma 2-ren gainean 4,3 milioi dokumentu eta 4.200 milioi token dituen euskarazko corpus batekin. Euskarazko kalitate handiko ebaluazio multzoen urritasunari aurre egiteko, lau ebaluazio multzo berri bildu ditugu: Eus- Proficiency, EGA azterketaren atariko probako 5.169 galdera; EusReading, irakurketaren ulermenerako 52 galdera; EusTrivia, ezagutza orokorreko 5 arlotako 1.715 galdera; eta EusExams, oposizioetako 16.774 galdera. Datu multzo berri horiekin ebaluatu ditugu Latxa eta euskarazko beste HEak (elebakar zein eleaniztunak), eta Latxak aurreko eredu ireki guztiak gainditu ditu. Halaber, GPT-4 Turbo HE komertzialarekin alderatuta, emaitza konpetitiboak lortu ditu hizkuntza-ezagutzan eta ulermenean, nahiz eta testu-irakurmenean zein ezagutza intentsiboa eskatzen duten atazetan atzeratuta egon. Latxa ereduak, corpusak eta ebaluazio-datu berriak lizentzia irekien pean daude eskuragarri, https://github.com/hitz-zentroa/latxa helbidean.

HITZ GAKOAK: Hizkuntzaren Prozesamendua, hizkuntza-ereduak, baliabide urriko hizkuntzak, euskara.

ABSTRACT: *We introduce the Latxa family of Large Language Models (LLMs), currently the largest developed for Basque. Latxa models range from 7 to 70 billion parameters and are built on LLama 2 models, which we continued pretraining on 4.3 million documents and 4.2 billion tokens of Basque. To address the scarcity of high-quality evaluation benchmarks for Basque, we collected four new datasets: EusProficiency, 5,169 Atarikoa test questions of EGA exams; EusReading, 352 reading comprehension questions; EusTrivia, with 1,715 general knowledge questions across 5 areas; and EusExams, 16,774 questions from public office exams. We conducted evaluations of Latxa and other LLMs (both monolingual and multilingual), with results showing Latxa's superiority over previous open models. It also obtains competitive results with the commercial GPT-4 Turbo in language proficiency and understanding, despite lagging behind in reading comprehension and knowledge-intensive tasks. Both the Latxa models, and our pretraining and evaluation data are publicly available under open licenses.*

KEYWORDS: Natural Language Processing, language models, low-resource languages, Basque.

*** *Harremanetan jartzeko* / *Corresponding author:*** Naiara Pérez. HiTZ Hizkuntza Teknologiako Euskal Zentroa-Ixa, Euskal Herriko Unibertsitatea (UPV/EHU), Sarriena auzoa, z/g (48091-Leioa, Euskal Herria). – naiara.perez@ehu.eus – https://orcid.org/0000-0001-8648-0428

Nola aipatu / ***How to cite:*** Pérez, Naiara; Etxaniz, Julen; Sainz, Óscar; Aldabe, Itziar; Rigau, German; Agirre, Eneko; Salem, Ahmed; Ormazabal, Aitor; Artetxe, Mikel; Soroa, Aitor (2025). «Latxa euskarazko hizkuntza-eredua». *Ekaia*, ale berezia 2025, 13-27. (https://doi.org/10.1387/ekaia.26338).

Jasotze-data: ; Onartze-data: .

ISSN 0214-9001 - eISSN 2444-3255 / © UPV/EHU Press 2025

Lan hau *Creative Commons Aitortu-EzKomertziala-PartekatuBerdin 4.0 Nazioartekoa* lizentzia baten mende dago

1. SARRERA

Hizkuntza-ereduek (HE) izan duten arrakastak interes komertzial izugarria piztu du azken boladan.[1] Zoritxarrez, egun dauden HE hoberenak konpainia pribatuek garatuak dira, hala nola GPT [2], Claude [3] edo Gemini [4]. Eredu irekiak sortzeko ekimenak sortu badira ere [5, 6, 7, 8], emaitza hoberenak lortzen dituzten HEek ingelesa eta ingelesezko kultura dituzte ardatz nagusi. 4. taulan arreta jartzea besterik ez dago HE irekiek euskaraz lortzen dituzten emaitza makalak ikusteko, askotan ausazko langatik oso gertu.

Artikulu honetan, Latxa aurkezten dugu, euskararako HE familia irekia, aurreko eredu irekiak baino emaitza aski hobeak lortzen dituena. Ezagun da HEak eraikitzeko testu-masa izugarri handiak behar direla. Adibidez, LLama 2 ereduak entrenatzeko, 2 bilioi hitz behar izan ziren [7], eta azken LLama 3 ereduak, berriz, 15 bilioi hitz behar izan ditu.[2] Horrek arazo larria dakar hizkuntza gutxituentzako HEak eraikitzeko garaian —euskaraz, kasurako. Izan ere, munduan hitz egiten diren hizkuntza gehienetan ez dago HEak eraikitzeko behar bezain beste testu atzigarri.

[9] lanean, finlandierazko HEak eraikitzeko hainbat estrategia probatu dituzte autoreek, eta erakutsi dute aurreikasketa jarraitua (*continual pretraining*) deritzon hurbilpena oso egokia dela baliabide urriko hizkuntzen HEak garatzeko. Eredua zerotik irakatsi ordez, aurretik ikasitako HEa oinarri hartzen da hurbilpen honetan, eta entrenamendu-prozesu bat egiten da, xede-hizkuntzako testua erabiliz. Latxa ereduak garatzeko, estrategia bera erabili dugu guk. Zehazki, LLama 2 HEan oinarritzen dira gure ereduak [7], zeina euskaraz entrenatzen jarraitu baitugu, 4,3 milioi dokumentu eta 4.200 milioi hitz duen euskarazko corpusa erabiliz.

Horretaz gain, euskarazko HEak ebaluatzeko datu sorten egungo gabeziari aurre egin diogu. Horrela, HEek euskaraz duten trebezia ebaluatzeko, aukera anitzeko lau datu multzo sortu ditugu ebaluaziorako, 23.282 galdera denera, eta publikoki banatu. Ebaluazio-multzoek hizkuntza-gaitasuna, irakurketaren ulermena, ezagutza orokorra eta ezagutza profesionala jorratzen dituzte.

Artikulu honetan, Latxa HEen familia eraikitzeko egindako urratsak azalduko ditugu. Lehendabizi, entrenamendurako erabili dugun corpusa deskribatuko dugu (2. atala), eta, gero, entrenamenduaren xehetasunak azalduko ditugu, 3. atalean. Latxa ereduak ebaluatzeko bildu dugun eba-

[1] Hizkuntza-ereduak zer diren eta zertarako erabiltzen diren jakiteko, ikusi, adibidez, [1] lana.

[2] https://ai.meta.com/blog/meta-llama-3/

luazio-datuen sortaz mintzatuko gara ondoren (4. atala), eta ebaluazioa bera nola egin dugun azalduko dugu (5. atala). Ebaluazio horren emaitzak 6. atalean erakutsiko ditugu, eta Latxa artearen egoerako beste HEekin alderatu. Azkenik, ondorioak azaldu eta etorkizuneko lanaz mintzatuko gara (7. atala).

2. ENTRENAMENDU CORPUSAK

Entrenatzeko erabilitako corpusa hainbat iturritatik eratorria da. Haietariko batzuk aurretik existitzen ziren, eta beste batzuk, berariaz bildu ditugu Latxa entrenatzeko. Corpusa biltzean, kalitatea izan dugu irizpide nagusi. Horretarako, kalitatezko iturriak aukeratu ditugu, eta garbiketak zein bikoizketak hautemateko aurreprozesu sendoak egin. Hurrengo ataletan, datu-iturriak azalduko ditugu (2.1. atala), corpusaren aurreprozesaketarekin batera (2.2. atala). Azken corpusaren estatistikak 1. taulan daude laburtuak.

2.1. Datu iturriak

EusCrawl v1.1. Euscrawl kalitate handiko corpusa da, bereziki aukeratutako 33 iturritatik eratorria. Izan ere, frogatu da iturriak eskuz aukeratzean lortutako testuaren kalitatea beste hurbilpenak eginez baino hobea dela [10]. Lan honetan EusCrawl hedatu dugu, metodologia berari jarraikiz, datuak 2023ko azarora arte hedatuta. Bertsio hori 384 milioi hitzez (1,94 milioi dokumentu) osatua dago.

1. taula. Datu-iturriak eta estatistikak aurreprozesaketaren urrats bakoitzean (Tok: LLama 2-ko tokenak).

	Gordin		Bikoizk gabe		Iragazia			Iturria
	Dok	Hitz	Dok	Hitz	Dok	Hitz	Tok	
CulturaX	1,60M	622M	1,33M	548M	1,31M	541M	1.840M	hf.co/uonlp/CulturaX
EusCrawl v1.1	2,12M	411M	1,94M	384M	1,79M	359M	1.210M	Artetxe *et al.* [10]
HPLT v1	2,29M	1.547M	1,56M	312M	0,37M	120M	421M	hplt-project.org
Colossal OSCAR	0,65M	283M	0,25M	111M	0,24M	105M	380M	hf.co/oscar-corpus
Wikipedia	0,55M	54M	0,55M	54M	0,41M	51M	182M	dumps.wikimedia.org
Egunkaria	0,18M	40M	0,18M	39M	0,18M	39M	129M	n/a
Booktegi	181	3M	177	3M	166	3M	8M	booktegi.eus
Guztira	7,39M	2.960M	5,80M	1.451M	4,30M	1.218M	4.170M	

Egunkaria. ***Egunkaria***ko edukia biltzen duen corpusa da, 176.000 berriz osatua eta 40 milioi hitz dituena.

Booktegi. Booktegi plataformako euskarazko eduki libreaz osatua dago, hala nola liburuak, elkarrizketak eta audio materialak. datu multzo honek 166 liburu elektroniko (3 milioi hitz) biltzen ditu: fikzioa, olerkiak eta saiakerak, besteak beste.

Wikipedia. Euskarazko Wikipedia jaitsi eta prozesatu dugu,[3] eta, hala, 550.000 dokumentu eta 54 milioi hitz lortu ditugu.

CulturaX. CulturaX [11] datu multzo elaniztuna da, mC4 [12] eta OSCARren 2019, 21.09, 22.01 eta 23.01 bertsioak konbinatuz osatua. 66 Common Crawl (CC) multzoko 2015etik 2022ra arteko bertsioez osatua dago [13]. CulturaX corpusean, % 2a da euskarazko zatia, eta 1,60 milioi dokumentuz (622 milioi hitz) osatuta dago.

Colossal OSCAR. Gaurdaino OSCAR proiektuko argitaratze handiena [13], Colossal OSCAR, CCeko 10 bertsioetatik eratorria da. Lan honetan, CulturaXk estaltzen ez dituen bi bertsio erabili ditugu, hots, 06-07-22 eta 05-06-23 bertsioak. Horretaz gain, 2023ko apirileko OSCAR bertsio bat ere erabili dugu, jadanik aurreprozesatua zegoena. Guztira, 648 mila dokumentu eta 283 milioi hitz biltzen ditu datu multzo honek.

HPLT v1. The High Performance Language Technologies ekimenak [14] beste datu multzo eleaniztun eta masibo bat bildu du, Internet Archive eta CCtik eratorria. Lan honetan, HLPTren lehenengo bertsioa erabili dugu, 2,29 milioi dokumentu dituena (1.550 milioi hitz). Aurreko datu multzoak ez bezala, HPLT ez dator aurreprozesatua, eta dokumentu errepikatuak ditu. Hori horrela, aurreprozesaketa sakona egin da datu multzo honen gainean, hurrengo atalean azalduko dugun bezala.

2.2. **Aurreprozesaketa**

Aurreko atalean azaldutako datu-iturriak berariaz aukeratu ditugu kalitatezko edukiak dituztelako, baina, hala ere, aurreprozesaketarako urrats bat egin dugu datuen gainean. Batez ere, gera zitekeen kalitate txarreko edukia ezabatu nahi izan dugu, eta baita datu sorten arteko bikoizketak ere; alegia, datu multzo bat baino gehiagotan azaltzen diren dokumentuak identifikatu eta kopia bakar batekin geratu gara. Aurreprozesaketa egiteko, Dolma toolkit [15] eta Corpus Cleaner v2 [16] tresnak erabili ditugu, zeinen bitartez dokumentuak normalizatu, iragazi eta bikoizketak identifikatu ditugun:

Normalizazioa. CCv2 erabili dugu karaktereen kodeketa eta zuriuneak normalizatzeko.

[3] 20231101 bertsioa, 2023ko azarokoa.

Bikoizketak identifikatu. Datu multzoen arteko bikoizketak identifikatu eta ezabatu ditugu, bai dokumentu mailan (dokumentuek helbide bera dutenean), eta baita edukiari erreparatuz ere. Azken hori egiteko Bloom iragazkiak [17] erabili ditugu, Dolma ingurunean inplementatuta daudenak. Corpusaren kalitatea bermatzeko, testu-edukia jatorriaren arabera lehenetsi dugu: lehenbizi, iturri fidagarrietatik eratorritako edukia (Wikipedia, EusCrawl, Egunkaria eta Booktegi); gero, CulturaX eta Colossal OSCAR datu-iturriak, eta, azkenik, HPLT. Azken horretan, bikoizketak paragrafo mailan identifikatu eta ezabatu dira.

Iragazketa. Bi urrats egin ditugu, kalitate txarreko edukia identifikatu eta ezabatzeko. Lehenbizi, kalitate txarreko testua identifikatzeko, Gopher [18] eta C4 [19] hizkuntza-ereduak eraikitzeko erabili ziren heuristikoak erabili ditugu, euskarara moldatuak (adibidez, batez besteko hitz luzera). Bestalde, euskaraz idatzita ez zeuden dokumentuak ezabatu ditugu, horretarako hizkuntza-ezagutzaile automatikoa erabiliz. Azken urrats horrek, HPLT datu multzoan izan du eragina batez ere, eta corpusaren herena baztertu dugu. Azkenik, testua CCv2 tresnarekin prozesatu dugu, zeinak dokumentu bakoitzari kalitate-kalifikazioa esleitzen dion. Berriro ere, HPLTn izan zuen eraginik handiena, eta, azken urratsaren ondoren, datu multzoa % 25 txikitu zen.

Urrats horien ondoko corpusaren tamainak eta estatistikak 1. taulan ikus daitezke. Guztira 1.220 milioi hitzez osatuta dago, eta LLama 2-ko 4.170 milioi token ditu. Corpusa ausaz nahastu, eta hiru zatitan banatu dugu: ikasteko (% 98), baliozkotzeko (% 1) eta ebaluatzeko (% 1).

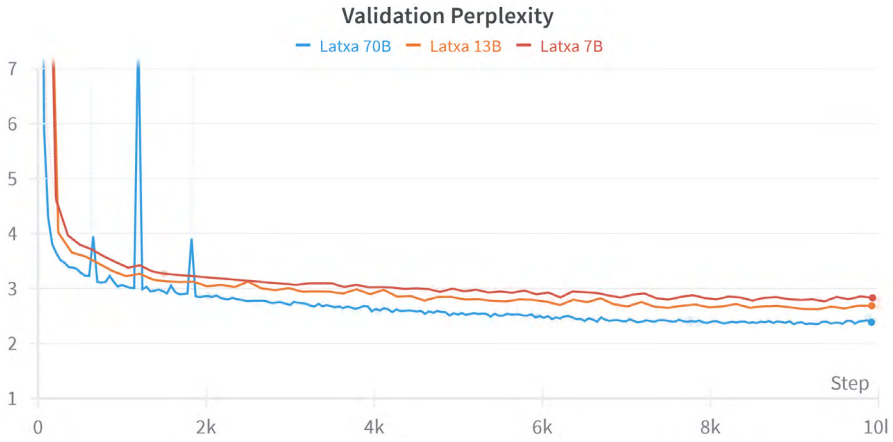
3. LATXA ENTRENATZEN

Hiru Latxa eredu eraiki ditugu, 7B, 13B eta 70B parametro dauzkatenak, hurrenez hurren.[4] Arestian aipatu bezala, horretarako LLama 2 HEa [7] hartu dugu abiapuntu, eta aurreikasketa jarraitua hurbilpena balitu dugu, hots, eredua entrenatzen jarraitu dugu 2.1. atalean azaldutako euskarazko corpusa erabiliz. Ahazte hondagarria (*catastrophic forgetting*) izeneko fenomenoa arintzeko —hots, eredua euskaraz irakasteagatik ingelesez zekiena ahaztea—, euskarazko corpusarekin batera ingelesezko testuarekin ere elikatu dugu Latxa. Horretarako, The Pile [20] multzoko 500.000 dokumentu aukeratu genituen ausaz, 900 milioi token guztira.

Latxa entrenatzeko, GPT-Neox [21] liburutegiaz baliatu gara. Prozesu guztia CINECA Leonardo Booster superkonputagailuan egin dugu,

[4] «B» sinboloak bilioi estatubatuarra adierazten du (gure mila miloia). Horrela, 7B ereduak 7.000 milioi parametro ditu, eta 70B ereduak, berriz, 70.000 milioi.

zeinak 3.456 nodo dituen, bakoitza lau A100 motako GPUrekin hornitua (64GB). Ereduak entrenatzeko, 10.000 urrats egin ditugu, 4.096 tokeneko sekuentziaren luzera erabili dugu, eta loteen tamaina 1 milioi tokeneko izan da. Guztira, 10.000 milioi token erabiliz entrenatu ditugu ereduak. Kosinuaren planifikatzailea erabili dugu ikaskuntza-tasa gidatzeko, 500 urratseko beroketa egin eta ikaskuntza-tasaren goreneko mugaren % 3ra arte jaitsiz (goreneko muga 1×10^{-4} balioan ezarri genuen). Gainontzeko hiperparametro guztiak [7] lanean azaltzen direnak dira. 1. irudian ikus dezakegu entrenamenduan zehar lortutako baliozkotze-harridura (*perplexity*).



1. irudia. Baliozkotze-harridura (*perplexity*) entrenamenduan zehar.

4. EBALUAZIORAKO DATU MULTZOAK

HEak ebaluatzeko egokiak diren euskarazko datu multzoen eskasia gainditzeko, ebaluazio-datu berriak bildu ditugu sareko hainbat joko eta probatatik. Hurbilpen horrek bi abantaila nagusi ditu beste hizkuntza batzuetako datu multzoen itzulpenarekin konparatuta: batetik, itzulpen artefaktuak [22] saihesten ditugu, eta, bestetik, HEek euskal hiztunentzat gertukoak edota garrantzitsuak diren gaiei buruz duten ezagutza ebalua dezakegu. Horrela, 4 datu multzo berri sortu ditugu. Bakoitzak HEen gaitasun edo ezagutza-arlo bat ebaluatzeko balio du, aukera anitzeko galderen bitartez. Jarraian, xehetasun handiagoz deskribatuko ditugu datu multzo horiek. Adibideak eta estatistikak 2. eta 3. tauletan jaso ditugu, hurrenez hurren.

EusProficiency. 1998tik 2008ra arteko EGA azterketetako atariko probetako 5.169 ariketak osatzen dute EusProficiency. Atarikoa EGAko

lehen proba puntuagarria zen, eta hizkuntza-gaitasunaren hainbat alderdi neurtzen zituen, hala nola irakurmena, gramatika, hiztegia, ortografia eta idazketa. Oro har, atariko test bakoitzak aukera anitzeko 85 galdera ditu, bakoitza lau aukera eta erantzun zuzen bakarraz osatuta.

EusReading. EGA azterketako irakurmeneko 352 probak osatzen dute EusReading. Proba bakoitzak aukera anitzeko 10 galdera izaten ditu, eta galdera bakoitzak lau aukera eta erantzun zuzen bakarra ditu. EusReading baliagarria da HEetann testu luzeen irakurketa- eta ulermen-gaitasuna neurtzeko.

EusTrivia. Sareko hainbat iturritatik hartutako ezagutza orokorrari buruzko 1.715 galderaz osatuta dago EusTrivia. Galderen % 56,3 Lehen Hezkuntzakoak dira, eta gainerakoak, berriz, konplexutzat hartzen dira. Galderen zati handi bat Euskal Herriari, euskal kulturari edo euskarari buruzkoa da. Aukera anitzeko galdera bakoitzak bi, hiru edo lau aukera ditu (3,84 batez beste) eta erantzun zuzen bakarra. Galderak bost jakintza-arlotan daude banatuta:

— **Gizarte- eta natur-zientziak** (% 27,8): historia, geografia, biologia, ekologia eta beste gizarte- eta natur-zientziei buruzko galderak.
— **Aisia eta artea** (% 24,5): kirolak eta kirolariak, arte performatiboak eta plastikoak, artistak, arkitektura, kultura-ekitaldiak eta antzeko gaiak.
— **Musika** (% 16,0): musika klasikoari zein garaikideari eta musikariei buruzko galderak.
— **Hizkuntza eta literatura** (% 17,1): mota guztietako literatura-ekoizpenei eta idazleei buruzko galderak, bai eta gai metalinguistikoak ere (definizioak, sinonimoak eta abar).
— **Matematika eta IKTak** (% 14,5): problema matematikoak eta IKTei buruzko galderak, edota arlo horietan egindako ekarpenengatik ezagunak diren pertsonei buruzko galderak.

EusExams. Euskal Herriko hainbat erakundek, hala nola Osakidetzak, Eusko Jaurlaritzak, Bilboko eta Gasteizko Udalek eta Euskal Herriko Unibertsitateak egindako oposizio-azterketetarako prestatzeko testen bilduma da EusExams. Lanpostu publiko askotarako azterketak biltzen ditu; adibidez, administrari- eta laguntzaile-lanetarako oposizioak, bai eta zeladore-, erizain- edota sukaldari-lanetarakoak ere. Hau da, maila profesionaleko ezagutza ebaluatzeko balio du EusExamsek. Denera, 16.774 galderak osatzen dute datu multzo hau.

2. taula. Ebaluaziorako datu multzoen adibideak.

EusProficiency

Galdera: Jatetxe batera sartu, eta bazkaltzen ari denari:

A. Gabon!
B. On egin diezazula!
C. Bejondeizula!
D. Agur t'erdi!

Erantzuna: B

EusReading

Pasartea: Ernest Hemingway, berak jakin barik, azkenekoz etorri da Bilbora, eta oro har, Penintsulara. Eta hori tamala, hilak 24 dituelarik Bilbon zezenak Ordoñez bere kutunari adarkada ederra sartu dio. Ez da ezer izan, zorionez. Biharamuneko *El Correo Español* egunkarian emandako argazkian ikusten den legez, idazleak bisita egin dion unean, toreatzailea hortxe dago, ondo bizirik, ohean. [...]

Galdera: 1960ko abuztuaren 24an.

A. *El Correo Español*-eko C. Barrenarekin batera agertzen den Ordoñez toreatzailea harrapatu zuen zezen batek.
B. Ernest Hemingwayk Bilboko plazan adarkada jaso zuen Ordoñez toreatzaileari bisita egin zion.
C. *El Correo*-ko argazkian, zezen batek Ordoñez toreatzailea harrapatzen du.
D. Ernest Hemingway lehenengo eta azkeneko aldiz iritsi zen Bilbora.

Erantzuna: B.

EusTrivia

Galdera: Zenbat kilo dauka tona batek?

A. 10.000 kilo.
B. 1.000.000 kilo.
C. 1.000 kilo.
D. 100 kilo.

Erantzuna: C.

EusExams

Galdera: UPV/EHUREN ONDAREA HAU DA:

A. UPV/EHUk jabetzan dituen ondasunak.
B. UPV/EHUk jabetzan dituen ondasun eta eskubideak.
C. UPV/EHUk jabetzan edo titularitatean dituen ondasun eta eskubideak, bai eta etorkizunean eskuratzen edo esleitzen zaizkion gainerako guztiak ere.
D. UPV/EHUk jabetzan dituen ondasunak, bai eta etorkizunean eskuratzen dituen gainerako guztiak ere.

Erantzuna: C.

3. taula. Ebaluazio-multzoen estatistikak: adibide kopurua, sarreraren eta irteeraren batez besteko luzera karakteretan, eta aukera kopurua adibide bakoitzeko.

	Sarrera		Irteera	
	Adibide	Karaktere	Aukera	Karaktere
EusProficiency	5.169	50	4	28
EusReading	352	5.340	2-4	67
EusTrivia	1.715	55	2-4	14
EusExams	16.774	115	4	63

5. **ESPERIMENTAZIO-INGURUNEA**

Latxa ereduen kalitatea neurtu dugu, eta beste hizkuntza-eredu batzuekin alderatu. Batetik, LLama 2 ereduarekin [7] alderatu dugu, aurreikasketa jarraituaren prozesuaren arrakasta neurtzeko balio digun neurrian. Horretaz gain, artearen egoerako hainbat HErekin alderatu dugu, bai euskaraz ez dakiten eredu eleaniztunekin (Mistral-7B [23], Mixtral-8x7B [8], 01.AI's Yi-34B [24]), eta baita euskaraz badakitenekin ere (BLOOM-7B [6], XGLM-7B [5] eta mGPT-13B [25]). Azkenik, HE komertzialekin ere alderatu nahi izan dugu, hots, OpenAI enpresako GPT-3.5 Turbo (gpt-3.5-turbo-0125) eta GPT-4 Turbo (gpt-4-1106-preview) ereduekin.

Ereduak *LM Evaluation Harness* izeneko liburutegia erabiliz ebaluatu ditugu [26]. Gure ebaluaziorako datu multzo guztiak aukera anitzeko adibideak dira, hots, ariketa azaltzen duen testuaren ondoren, lau aukera eskaintzen dira, eta ataza ebazteko aukera egokia aukeratu behar da (A, B, C edo D; ikusi adibideak 2. taulan). HEak ebaluatzeko, lau testuak eskaintzen zaizkio sistemari (testuingurua, gehi aukera posible bat), eta HEak aukera bakoitzari ematen dion probabilitatea kalkulatzen da. Probabilitate handiena lortzen duena izango da HEak aukeratutakoa. Horretaz gain, *5-shot learning* izeneko hurbilpena egin dugu; alegia, galderaren aurretik, ebatzitako 5 adibide ematen zaizkio sistemari, HEak uler dezan ataza zertan den. GPT eredu komertzialak frogatzeko, haren APIa erabili dugu, baita esperimentazio-ingurune bera ere.

6. **EMAITZAK**

Emaitza nagusiak 4. taulan ikus daitezke. Oro har, Latxak emaitzarik hoberenak lortzen ditu beste HE irekiekin konparatzen badugu, eta alde handia ateratzen die: gure eredurik hoberenak batez beste 56,39 puntu lortu ditu, eta bigarren eredu irekiari 19,81 puntu ateratzen dizkio. Nabaria da ere tamainaren garrantzia. Izan ere, HEak zenbat eta parametro gehiago

eduki, emaitza ere orduan eta hobea da. Horrela, emaitzarik hoberenak 70B ereduak lortzen ditu. Nabarmentzekoa da hain modelo handiak ere molda daitezkeela euskarara, nahiz eta aurreikasketa jarraituan erabilitako datu kopurua konparatiboki urria izan (10.000 milioi token). Emaitzak iradokitzen du Latxak lortutako emaitzak hobetu litezkeela etorkizunean, oinarri gisa HE hobeagoa bat bada abiapuntu.

Latxak GPT3.5 Turbok baino emaitza hobeak lortu ditu datu multzo guztietan. GPT4 Turbo, berriz, Latxa baino hobeagoa da, baina ez *EusProficiency* datu multzoan, hots, euskararen gaitasuna neurtzen duen multzoan. Alegia, badirudi atazak ebazteko gaitasuna ez dagoela soilik baldintzatuta HEak hizkuntza horretan duen gaitasunean.

4. taula. Emaitza nagusiak. Klase bakoitzeko emaitza hoberenak **beltzez** daude adierazita. Emaitza orokor hoberenak, berriz, azpimarratuak daude.

		EusProf	EusRead	EusTrivia	EusExams	Batez beste
Ausaz		25,00	25,83	26,55	25,00	25,60
GPT-3,5 Turbo	n/a	31,24	36,65	46,71	42,42	39,26
GPT-4 Turbo	n/a	**56,70**	**75,85**	**73,12**	**70,22**	**68,97**
XGLM	7B	22,96	24,43	26,53	24,59	24,63
BLOOM	7B	25,34	28,41	27,17	25,07	26,50
Mistral	7B	25,01	**29,26**	34,58	32,15	30,25
LLama 2	7B	24,09	27,27	29,50	28,84	27,43
Latxa	7B	**30,26**	25,00	**42,16**	**33,82**	**32,81**
mGPT	13B	25,00	24,15	27,17	25,73	25,51
LLama 2	13B	25,90	28,98	33,53	29,66	29,52
Latxa	13B	**44,11**	**32,67**	**56,38**	**43,66**	**44,21**
Mixtral	8×7B	26,43	37,50	42,51	39,87	36,58
Yi	34B	27,30	34,66	42,57	39,68	36,05
LLama 2	70B	24,16	27,84	38,43	33,08	30,88
Latxa	70B	**60,65**	**50,57**	**62,45**	**51,90**	**56,39**

7. ONDORIOAK ETA ETORKIZUNEKO LANA

Lan honetan, Latxa aurkeztu dugu, euskararako hizkuntza-eredu (HE) sortzaileen familia, 7B, 13B eta 70B parametroko kopuruak dituztenak. Latxa HEa LLama 2-n oinarrituta dago, zeinaren gainean ikasketa jarraitua egin dugun, berariaz bildutako euskarazko corpusa erabiliz. Bildutako corpus hori da, egun, euskararako publikoki atzigarri dagoen handiena. Aurreprozesaketa egin eta bikoizketak ezabatu ondoren, corpusak 1.218 milioi hitz eta 4.170 milioi token biltzen ditu. Latxa ebaluatzeko, lau datu multzo

berri bildu ditugu, HEen euskara-gaitasuna zein euskal kulturaren ezagutza neurtzeko egun dagoen ebaluazio sortarik handiena.

Esperimentuek frogatzen dute Latxa beste HE irekiak baino nabarmen hobeagoa dela euskarazko datu multzoan, alde handia ateratzen baitie, eta baita ere GPT 3.5 eredu komertziala baino hobeagoa ere, nahiz eta kasu horretan aldea hain handia ez izan. GPT-4 Turbo baino okerragoa da Latxa datu multzo gehienetan, baina, euskarazko gaitasuna neurtzen duen ebaluazio-multzoan, —EusProficiencyn, alegia—, GPT-4 Turbok baino emaitza hobeagoa ematen du Latxak. Emaitzak iradokitzen du HEak ez direla soilik hizkuntzaren ezagutzan oinarritzen ataza konplexuak ebazteko garaian.

Etorkizunean, gure asmoa honako hau da: entrenamendu-corpusa gehiago handitzea, euskarazko iturri mota askotatik, hala nola argitalpenetatik eta egunkarietatik. Ebaluazio multzoa ere hedatu nahi dugu, egiazkotasuna edo haluzinazioak neurtzeko balio dezaten. Azkenik, Latxa entrenatzen jarraitu nahi dugu, giza aginduak bete ditzan.

ESKERRAK

Lan honek Eusko Jaurlaritzaren babesa jaso du (Errendimendu altuko taldea IT-1805-22 eta IKER-GAITU proiektua), eta baita ILENIA proiektuarena (2022/TL22/00215335), Eraldaketa Digitalerako eta Funtzio Publikorako Ministerioarena eta NextGenerationEU Recovery erakundeak finantzatutako eraldaketa- eta erresilientzia-planarena. Ereduak CINECAko Leonardo superkonputagailuan eraikiak izan dira, EuroHPC Joint Undertaking erakundeko EHPC-EXT-2023E01-013 proiektuaren babesean. Julen Etxanizek eta Oscar Sainzek Eusko Jaurlaritzak finantzatutako tesia egiteko beka eskuratu dute (PRE_2023_2_0060 eta PRE_2023_2_0137, hurrenez hurren).

ERREFERENTZIAK

[1] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie eta J.-R. Wen, 2023, «A survey of large language models», *arXiv preprint arXiv:230318223*.

[2] OpenAI, 2023, «GPT-4 technical report», *arXiv preprint arXiv:230308774*.

[3] S. Wu, M. Koo, L. Blum, A. Black, L. Kao, F. Scalzo eta I. Kurtz, 2023, «A comparative study of open-source large language models, GPT-4 and Claude 2: Multiple- choice test taking in nephrology», *arXiv preprint arXiv:230804709*.

[4] Gemini Team, 2023, «Gemini: A family of highly capable multimodal models», *arXiv preprint arXiv:231211805*.

[5] X. V. Lin, T. Mihaylov, M. Artetxe, T. Wang, S. Chen, D. Simig, M. Ott, N. Goyal, S. Bhosale, J. Du, R. Pasunuru, S. Shleifer, P. S. Koura, V. Chaudhary, B. O'Horo, J. Wang, L. Zettlemoyer, Z. Kozareva, M. Diab, V. Stoyanov eta X. Li, 2022, «Few-shot learning with multilingual generative language models», Y. Goldberg, Z. Kozareva eta Y. Zhang, editoreak, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Orrialdeak 9019-9052, Association for Computational Linguistics, Abu Dhabi, Arabiar Emirerri Batuak.

[6] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé eta kolaboratzaileak, 2023, «BLOOM: A 176B-parameter open-access multilingual language model», *arXiv preprint arXiv:221105100*.

[7] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale eta kolaboratzaileak, 2023, «Llama 2: Open foundation and fine-tuned chat models», *arXiv preprint arXiv:230709288*.

[8] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix eta W. E. Sayed, 2024, «Mixtral of experts», *arXiv preprint arXiv:240104088*.

[9] R. Luukkonen, V. Komulainen, J. Luoma, A. Eskelinen, J. Kanerva, H.-M. Kupari, F. Ginter, V. Laippala, N. Muennighoff, A. Piktus, T. Wang, N. Tazi, T. Scao, T. Wolf, O. Suominen, S. Sairanen, M. Merioksa, J. Heinonen, A. Vahtola, S. Antao eta S. Pyysalo, 2023, «FinGPT: Large generative models for a small language», H. Bouamor, J. Pino eta K. Bali, editoreak, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Orrialdeak 2710-2726, Association for Computational Linguistics, Singapur.

[10] M. Artetxe, I. Aldabe, R. Agerri, O. Perez-de Viñaspre eta A. Soroa, 2022, «Does corpus quality really matter for low-resource languages?», Y. Goldberg, Z. Kozareva eta Y. Zhang, Editoreak, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Orrialdeak 7383-7390, Association for Computational Linguistics, Abu Dhabi, Arabiar Emirerri Batuak.

[11] T. Ñguyen, C. V. Nguyen, V. D. Lai, H. Man, N. T. Ngo, F. Dernoncourt, R. A. Rossi eta T. H. Nguyen, 2024, «CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages», N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti eta N. Xue, editoreak, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Orrialdeak 4226-4237, ELRA eta ICCL, Torino, Italia.

[12] L. XUE, N. CONSTANT, A. ROBERTS, M. KALE, R. AL-RFOU, A. SIDDHANT, A. BARUA eta C. RAFFEL, 2021, «Mt5: A Massively Multilingual Pre-Trained Text-To-Text Transformer», K. TOUTANOVA, A. RUMSHISKY, L. ZETTLEMOYER, D. HAKKANITUR, I. BELTAGY, S. BETHARD, R. COTTERELL, T. CHAKRABORTY eta Y. ZHOU, editoreak, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Orrialdeak 483-498, Association for Computational Linguistics, online.

[13] P. J. ORTIZ SUÁREZ, B. SAGOT eta L. ROMARY, 2019, «Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures», P. BAŃSKI, A. BARBARESI, H. BIBER, E. BREITENEDER, S. CLEMATIDE, M. KUPIETZ, H. LÜNGEN eta C. ILIADI, editoreak, *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, Orrialdeak 9-16, Leibniz-Institut für Deutsche Sprache, Mannheim.

[14] M. AULAMO, N. BOGOYCHEV, S. JI, G. ÑAIL, G. RAMÍREZ-SÁNCHEZ, J. TIEDEMANN, J. VANDERLINDE eta J. ZARAGOZA, 2023, «HPLT: High performance language technologies», M. ÑURMINEN, J. BRENNER, M. KOPONEN, S. LATOMAA, M. MIKHAILOV, F. SCHIERL, T. RANASINGHE, E. VANMASSENHOVE, S. A. VIDAL, N. ARANBERRI, M. ÑUNZIATINI, C. P. ESCARTÍN, M. FORCADA, M. POPOVIC, C. SCARTON eta H. MONIZ, Editoreak, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, Orrialdeak 517-518, European Association for Machine Translation, Tampere, Finlandia.

[15] L. SOLDAINI, R. KINNEY, A. BHAGIA, D. SCHWENK, D. ATKINSON, R. AUTHUR, B. BOGIN, K. CHANDU, J. DUMAS, Y. ELAZAR, V. HOFMANN, A. JHA, S. KUMAR, L. LUCY, X. LYU, N. LAMBERT, I. MAGNUSSON, J. MORRISON, N. MUENNIGHOFF, A. ÑAIK, C. ÑAM, M. PETERS, A. RAVICHANDER, K. RICHARDSON, Z. SHEN, E. STRUBELL, N. SUBRAMANI, O. TAFJORD, E. WALSH, L. ZETTLEMOYER, N. SMITH, H. HAJISHIRZI, I. BELTAGY, D. GROENEVELD, J. DODGE eta K. LO, 2024, «Dolma: an open corpus of three trillion tokens for language model pretraining research», L.-W. KU, A. MARTINS eta V. SRIKUMAR,,Editoreak, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Orrialdeak 15725-15788, Association for Computational Linguistics, Bangkok, Thailandia. URL: https://aclanthology.org/2024.acl-long.840

[16] J. PALOMAR-GINER, J. J. SAIZ, F. ESPUÑA, M. MINA, S. DADALT, J. LLOP, M. OSTENDORFF, P. ORTIZ SUAREZ, G. REHM, A. GONZALEZ-AGIRRE eta M. VILLEGAS, 2024, «A CURATEd CATalog: Rethinking the extraction of pretraining corpora for mid- resourced languages», N. CALZOLARI, M.-Y. KAN, V. HOSTE, A. LENCI, S. SAKTI eta N. XUE, editoreak, *Proceedings of the 2024 Joint International Conference on Computational Linguis-*

tics, Language Resources and Evaluation (LREC-COLING 2024), Orrialdeak 335-349, ELRA eta ICCL, Torino, Italia.

[17] B. H. BLOOM, 1970, «Space/time trade-offs in hash coding with allowable errors», *Commun ACM*, **13**(7), 422-426.

[18] J. W. RAE, S. BORGEAUD, T. CAI, K. MILLICAN, J. HOFFMANN, F. SONG, J. ASLANIDES, S. HENDERSON, R. RING, S. YOUNG, E. RUTHERFORD, T. HENNIGAN, J. MENICK, A. CASSIRER, R. POWELL, G. VANDENDRIESSCHE, L. A. HENDRICKS, M. RAUH, P.-S. HUANG, A. GLAESE, J. WELBL, S. DATHATHRI, S. HUANG, J. UESATO, J. MELLOR, I. HIGGINS, A. CRESWELL, N. MCALEESE, A. WU, E. ELSEN, S. JAYAKUMAR, E. BUCHATSKAYA, D. BUDDEN, E. SUTHERLAND, K. SIMONYAN, M. PAGANINI, L. SIFRE, L. MARTENS, X. L. LI, A. KUNCORO, A. ÑEMATZADEH, E. GRIBOVSKAYA, D. DONATO, A. LAZARIDOU, A. MENSCH, J.-B. LESPIAU, M. TSIMPOUKELLI, N. GRIGOREV, D. FRITZ, T. SOTTIAUX, M. PAJARSKAS, T. POHLEN, Z. GONG, D. TOYAMA, C. DE MASSON D'AUTUME, Y. LI, T. TERZI, V. MIKULIK, I. BABUSCHKIN, A. CLARK, D. DE LAS CASAS, A. GUY, C. JONES, J. BRADBURY, M. JOHNSON, B. HECHTMAN, L. WEIDINGER, I. GABRIEL, W. ISAAC, E. LOCKHART, S. OSINDERO, L. RIMELL, C. DYER, O. VINYALS, K. AYOUB, J. STANWAY, L. BENNETT, D. HASSABIS, K. KAVUKCUOGLU eta G. IRVING, 2022, «Scaling language models: Methods, analysis & insights from training Gopher», *arXiv preprint arXiv:211211446*.

[19] C. RAFFEL, N. SHAZEER, A. ROBERTS, K. LEE, S. ÑARANG, M. MATENA, Y. ZHOU, W. LI eta P. J. LIU, 2020, «Exploring the limits of transfer learning with a unified text-to-text transformer», *J Mach Learn Res*, **21**(1), 140:1-140:67.

[20] L. GAO, S. BIDERMAN, S. BLACK, L. GOLDING, T. HOPPE, C. FOSTER, J. PHANG, H. HE, A. THITE, N. ÑABESHIMA eta kolaboratzaileak, 2020, «The Pile: An 800GB dataset of diverse text for language modeling», *arXiv preprint arXiv:210100027*.

[21] A. ANDONIAN, Q. ANTHONY, S. BIDERMAN, S. BLACK, P. GALI, L. GAO, E. HALLAHAN, J. LEVY-KRAMER, C. LEAHY, L. ÑESTLER, K. PARKER, M. PIELER, J. PHANG, S. PU-ROHIT, H. SCHOELKOPF, D. STANDER, T. SONGZ, C. TIGGES, B. THÉRIEN, P. WANG eta S. WEINBACH, 2023, «GPT-NeoX: Large scale autoregressive language modeling in PyTorch».

[22] M. ARTETXE, G. LABAKA eta E. AGIRRE, 2020, «Translation artifacts in cross-lingual transfer learning», B. WEBBER, T. COHN, Y. HE eta Y. LIU, editoreak, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Orrialdeak 7674-7684, Association for Computational Linguistics, online.

[23] A. Q. JIANG, A. SABLAYROLLES, A. MENSCH, C. BAMFORD, D. S. CHAPLOT, D. DE LAS CASAS, F. BRESSAND, G. LENGYEL, G. LAMPLE, L. SAULNIER, L. R. LAVAUD, M.-A. LACHAUX, P. STOCK, T. L. SCAO, T. LAVRIL, T. WANG, T. LACROIX eta W. E. SAYED, 2023, «Mistral 7B», *arXiv preprint arXiv:231006825*.

[24] 01.AI, 2024, «Yi: Open foundation models by 01.AI», *arXiv preprint arXiv:240304652*.

[25] O. Shliazhko, A. Fenogenova, M. Tikhonova, A. Kozlova, V. Mikhailov eta T. Shavrina, 2024, «mGPT: Few-shot learners go multilingual», *Transactions of the Association for Computational Linguistics*, **12**, 58-79.

[26] S. Biderman, H. Schoelkopf, L. Sutawika, L. Gao, J. Tow, B. Abbasi, A. F. Aji, P. S. Ammanamanchi, S. Black, J. Clive, A. Dipofi, J. Etxaniz, B. Fattori, J. Z. Forde, C. Foster, J. Hsu, M. Jaiswal, W. Y. Lee, H. Li, C. Lovering, N. Muen-Nighoff, E. Pavlick, J. Phang, A. Skowron, S. Tan, X. Tang, K. A. Wang, G. I. Winata, F. Yvon eta A. Zou, 2024, «Lessons from the trenches on reproducible evaluation of language models», *arXiv preprint arXiv:240514782*.