IKER GAZTE
NAZIOARTEKO IKERKETA EUSKARAZ

VI. IKERGAZTE

NAZIOARTEKO IKERKETA EUSKARAZ

2025eko maiatzaren 28, 29 eta 30a
Bilbo, Euskal Herria

ANTOLATZAILEA:
Udako Euskal Unibertsitatea (UEU)

INGENIARITZA ETA ARKITEKTURA

Hizkuntzetarako neurona espezifikoak LLMetan?

Ixak Sarasua eta *Xabier Saralegi*

227-233 or.
https://dx.doi.org/10.26876/ikergazte.vi.03.28

Hizkuntzetarako neurona espezifikoak LLMetan?

Ixak Sarasua[1,2], Xabier Saralegi[1]

[1] Orai NLP Teknologiak. Zelai Haundi kalea, 3, Osinalde industrialdea, 20170 Usurbil, Gipuzkoa
[2] Euskal Herriko Unibertsitatea UPV/EHU

i.sarasua@orai.eus

Laburpena

Hizkuntza Eredu Handiak (LLM) adimen artifiziala irauli duten milaka milioi parametroko neurona-sareak dira. Ikerketa honetan LLMen hizkuntza jakinetarako neurona espezifikoak aztertzen dira, euskaran fokua ezarriz. Hizkuntza Aktibaziorako Probabilitate Entropia (LAPE) metrika erabiliz, Llama-3.1-8B ereduko eta euskarara egokitutako aldaerako (Llama-eus-8B) euskara, frantses, gaztelera eta ingeles hizkuntzetan espezializatutako neuronak identifikatzen ditugu. Esperimentuetan ikusten da neurona espezifikoak ereduaren kanpoaldeko geruzetan pilatzen direla gehienbat, eta euskarak dituela neurona espezifiko gehien. *Perplexity*-a erabiliz egindako analisiak erakusten du neurona horiek desaktibatzeak eragin berezia duela helburuko hizkuntzan ereduko hizkuntza nagusia ez den kasuetan, neuronen espezifikotasuna baieztatuz. Aurkikuntza horiek horrelako ereduak beste hizkuntzetara egokitzearen eta neurona espezializatuen arteko erlazioa erakusten dute, eta LLMak baliabide urriko hizkuntzetara era optimoan egokitzeko bideei buruzko informazioa ematen dute.

Hitz gakoak: LLM, euskara, baliabide urriko hizkuntza, interpretazio mekaniko, neurona espezifiko

Abstract

Large Language Models (LLM) are revolutionary artificial intelligence neural networks with billions of parameters. This study investigates language-specific neurons in LLMs, focusing on their adaptation to Basque. Using the Language Activation Probability Entropy (LAPE) metric, we identify neurons in Llama-3.1-8B and its Basque-adapted variant (Llama-eus-8B) that are specialized for Basque, French, Spanish, and English. Experiments reveal that language-specific neurons predominantly cluster in final layers, with Basque showing the highest count. Perplexity analysis indicates that deactivating these neurons disproportionately affects their target languages when these are not predominant in the model, validating their specificity. These findings highlight the interplay between language adaptation and neuron distribution, offering insights for optimizing LLMs for low-resource languages.

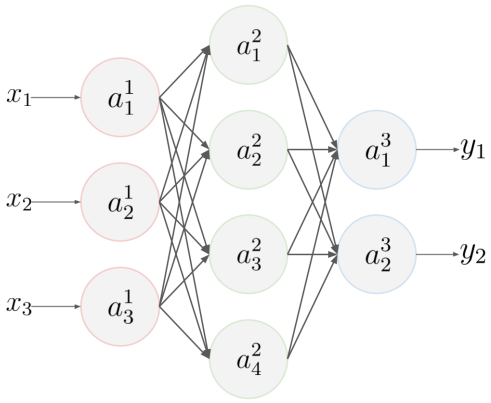
Keywords: *LLM, Basque, low-resourced language, mechanistic interpretation, specific neuron*

1 Sarrera eta motibazioa

Hizkuntza Eredu Handiek (ingelesez *Large Language Model*, aurrerantzean LLM) Hizkuntza Naturalaren Prozesamenduaren mundua irauli dute azken urteetan. Llama-3, o1, DeepSeek-R1 eta antzeko ereduek (Dubey *et al.*, 2024; OpenAI, 2024; Guo *et al.*, 2025) laburpen automatikoa, galdera-erantzuna, ideien sorkuntza eta, oro har, giza-hizkuntza ulertu eta sortzea behar duen edozein ataza egiten dute.

Izenak dioen bezala, LLMak tamaina handikoak dira: milaka milioi parametro izan ohi dituzte, eta bilioika hitzeko corpusekin entrenatzen dira. Entrenamendu horretan hainbat gaitasun lortzen dituzte: hizkuntzaren ortografia-, sintaxi- eta gramatika-arauen ezagutza, hitzen esanahi semantikoei buruzko jakinduria, baina baita munduari buruzko ezagutza eta hizkuntzarekin lotutako atazak egiteko eta arrazoitzeko gaitasuna ere (Zhang *et al.*, 2024; Chang *et al.*, 2024). Jakintza hori guztia LLMaren parametroetan dago kodetuta, nahiz eta oso zaila den kodeketa horren intepretazio mekanikoak egitea; hau da, ez dago batere argi gramatikarako gaitasuna zein geruzatan dagoen, edo zein parametro diren munduari buruzko ezagutza-arlo jakin batekin lotura estuena dutenak (Mishra *et al.*, 2023).

1. Irudia: Hiru geruzako sare neuronal artifizial sinplea. Sare honek hiru dimentsioko bektore bat jasotzen du eta bi dimentsioko bat itzultzen du.



Baliabide urriko hizkuntzen ikuspegitik, askoz ere corpus txikiagoak izanik, zerotik LLMa hizkuntza horretan entrenatzea baino eraginkorragoa da baliabide gehiagoko hizkuntza batean (ingelesez, esate baterako) erabiltzeko entrenatu den LLM bat egokitzea (Etxaniz *et al.*, 2024; Corral *et al.*, 2024; Nag *et al.*, 2024). Estrategia horren bidez, LLMak baliabide askoko hizkuntzan ikasitako gaitasunak (hizkuntza guztiek partekatzen duten egitura oinarrizkoa, munduari buruzko ezagutza, hizkuntzarekin lotutako atazak, etab.) mantentzea nahi da, baliabide urriko hizkuntza ikasi eta gaitasun horiek hizkuntza berri horretan ere txertatzearekin batera. Egokitzapen hori egiteko modu bat baino gehiago dago, eta oso ikerketa-arlo aktiboa da egokitzapenak egiteko teknikena (Li *et al.*, 2024; Yamaguchi *et al.*, 2024).

LLMak beste hizkuntza batzuetara moldatzeko teknikei ahalik eta etekinik handiena ateratze aldera, helburua aurretik lortutako gaitasunak mantendu eta baliabide urriko hizkuntzara transferitzea baldin bada, interesgarria da jakitea zein diren mantendu nahi den gaitasun hori gordetzen duten parametroak, eta zein hizkuntza berria LLMari txertatzeko aldatu behar direnak. Era horretan, egokitzapenerako entrenamenduak moldatzeko teknikak garatu ahal izango lirateke helburua lortzeko. Giza-burmuinarekin analogia eginez, ingeles hiztun bati euskara ulertu eta euskaraz adierazten ikasten laguntzeko neurokirurgiako ebakuntza bat egitea izango litzateke ideia.

Artikulu honetan egiten diren ekarpenak ondoengoak dira:

1. Ingelesez erabiltzeko LLM baten eta euskara egokitutako beste baten hainbat hizkuntzatarako neurona espezifikoak identifikatzen ditugu eta ereduen geruzetan nola banatzen diren aztertzen dugu.
2. Neurona espezifikoek hizkuntza sortzeko gaitasunean duten eragina aztertzen dugu.
3. Esperimentuen emaitzak ikusiz, neurona espezifikoen eta hizkuntza-egokitzapenaren inguruko interpretazio eta hipotesiak egiten ditugu.

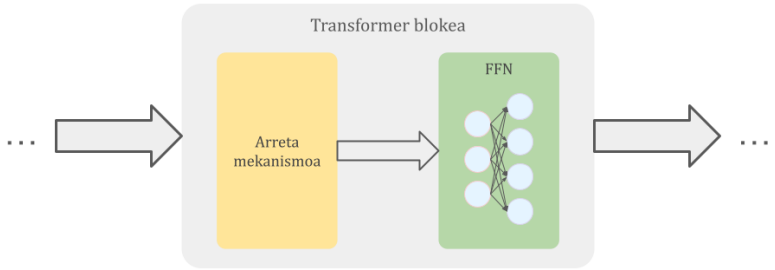
2 Arloko egoera eta ikerketaren helburuak

Sare neuronal artifizialak burmuinaren funtzionamenduan inspiratutako ereduak dira. Geruzatan antolatzen diren neurona deituriko unitate konektatuek osatzen dute sarea; neurona bakoitzak aurreko geruzetako neuronetatik seinale bana jasotzen du, seinale hori prozesatu eta hurrengo geruzara bidaltzen du. Neuronek hurrengo geruzara bidaltzen duten seinalea zenbaki erreal bat izaten da sare neuronal klasikoetan, 1 Irudiak erakusten duena bezalakoetan, eta zenbaki horri neuronaren aktibazio deitzen zaio. l-garren geruzako n neuronaren aktibazioa a_n^l idatziko dugu, eta honela kalkulatzen da:

$$a_n^l = \sigma\left(g_n^l\right), \tag{1}$$

$$g_n^l = \sum_{i=0}^{N_{l-1}} w_{i,n}^l a_i^{l-1} + b_n^l, \;\; l > 1. \tag{2}$$

2. Irudia: Transformer motako eredu baten bloke bakarraren irudi eskematikoa.



$w^l_{i,n}$ eta b^l_n neuronaren parametroak dira, lehenengoa $l-1$ geruzako i neuronaren aktibazioari dagokion pisua da eta bigarrena l-garren geruzako n neuronaren gai askea. N_{l-1} zenbakiak $l-1$ geruzan dauden neurona kopurua adierazten du, eta σ funtzio ez-lineal bat da, aktibazio-funtzio deritzona. Hainbat aktibazio-funtzio daude, eta ohikoenetako batzuk funtzio sigmoidea, tangente hiperbolikoa, ReLU funtzioa eta GeLU funtzioa dira (Dubey *et al.*, 2022). g^l_n-ren balioak atalase bat gainditzen baldin badu, neurona aktibatu egiten dela esaten da, eta, bestela, neurona inaktibo gelditzen da. Horren arrazoia zera da: g^l_n-ren balioa atalasetik behera baldin badago, σ funtzioak 0 balioa hartzen du, hau da, ez du seinalerik bidaltzen hurrengo geruzara: inaktibo geratzen da.

Hizkuntza Naturalaren Prozesamenduaren arloan (ingelesez Natural Language Processing, NLP) Transformer arkitektura (Vaswani, 2017) estandar bilakatu da azken urteetan. Transformer motako ereduak neurona-sareen orokorpena dira, datu sekuentzialak (adibidez, testuak) modelizatzeko diseinatuak. Arkitektura horren bereizgarri garrantzitsuena arreta mekanismoa da, zeinaren bidez ereduak prozesatzen ari den datu-unitateen (testuaren kasuan hitzak edo hitzen zatiak dira datu-unitateak) arteko erlazioak ikasten eta erabiltzen dituen. Arreta mekanismoan ez dago aktibazio funtziorik, eta, beraz, ezin da neuronen aktibazioa definitu mekanismo horretan.

Hala ere, Transformer motako ereduetan beste geruza mota bat ere badago, FFN geruza deiturikoa (*Feedforward Network*, Aurreranzko Elikadurako Sarea), 2 Irudian ikusten den bezala. Mota horretako geruzak (1) eta (2) ekuazioetan deskribatutakoen antzerakoak dira, eta paper garrantzitsua betetzen dute ereduak dituen gaitasunetan (Dai *et al.*, 2022). Llama-3.1 eredu-familian, esaterako, ereduen parametro guztien %70etik gora FFN geruzetan daude (Dubey *et al.*, 2024). FFN geruzetako neurona bereziak identifikatzeko metodo bat ikusiko dugu artikulu honetan, hizkuntza jakin baterako espezifikoak diren neuronak identifikatzekoa, hain zuzen.

Esperimentuetarako LLMa aukeratzeari dagokionez, Llama-3.1-8B (Dubey *et al.*, 2024) eta Llama-eus-8B (Corral *et al.*, 2024) ereduak hautatu ditugu. Llama-3.1 familiako ereduak, *Meta*k kaleratuak, kode irekikoak dira eta ingelesa dute hizkuntza nagusi, eta emaitza bikainak ematen dituzte hizkuntza horretan. Llama-eus-8B Llama-3.1-8B ereduaren euskararako egokitzapena da, *continual pretraining* teknika erabilita egokitu dena hain zuzen, hau da, ingeleserako LLMa euskarazko testuarekin entrenatzen jarraitu da hizkuntza horretako gaitasunak hobetzeko. Bi eredu horiek aztertuta, interesgarria da ikustea neurona espezifikoen konfigurazioa nola aldatzen den egokitzapen horrekin.

2.1 Hizkuntza Aktibaziorako Probabilitate Entropia

LLMen FFN geruzetako neuronetan jasota dagoen jakintza espezifikoa ikerketa-gai aktiboa da (Dai *et al.*, 2022; Lai *et al.*, 2024; Leng eta Xiong, 2024), eta hizkuntza bakoitzerako espezifikoak diren neuronak ereduan nola banatuta dauden jakiteko metodo bat ere badago (Tang *et al.*, 2024). Metodo horretan, Hizkuntza Aktibaziorako Probabilitate Entropia (*Language Activation Probability Entropy*, LAPE) deituriko metrika kalkulatzen da ereduko neurona bakoitzerako, eta LAPE txikia duten eta hizkuntza jakin batekin aktibatzeko probabilitate handia duten neuronak hizkuntza horretako espezifikotzat hartzen dira. Horrela kalkulatzen da hizkuntza bakoitzarekin LLMko neurona bakoitzak aktibatzeko duen probabilitatea:

$$p^k_{l,n} = \mathbb{E}\Big(\mathbb{I}\Big(\sigma\big(\textstyle\sum_i w^l_{i,n} a^{l-1}_i\big) > 0\Big) \mid \text{hizkuntza} = k\Big). \tag{3}$$

$p^k_{l,n}$ l geruzako n neuronak k hizkuntzarekin aktibatzeko duen probabilitatea da. $\mathbb{I}$ funtzio adierazlea da, zeinak 1 balioa hartzen duen barrukoa egia bada, eta 0 gezurra bada.

Horrela, $\boldsymbol{p}_{l,n} = (p^1_{l,n}, ..., p^K_{i,j})$ banaketa lortzen dugu neurona bakoitzarentzat. Hala ere, $\boldsymbol{p}_{l,n}$ ez da probabilitate-banaketa onargarria, orokorrean probabilitate guztien batura ez baita 1 izango; beraz, L1 normalizazioa aplikatzen da $\boldsymbol{p}'_{l,n}$ banaketa onargarria lortzeko. Azken banaketa horren entropiari deitzen zaio Hizkuntza Aktibaziorako Probabilitate Entropia, LAPE:

$$\text{LAPE}_{l,n} = -\sum_{k=1}^{K} p'^k_{i,j} \log(p'^k_{i,j}). \tag{4}$$

LAPE txikia duten neuronak hizkuntzaren baterako espezifikoak diren neuronak dira, aktibazio-probabilitatea nabarmen handiagoa izango baitu hizkuntzetako batean besteetan baino. Era horretan, LAPE txikiena duten neuronen $\%1$eko pertzentila hartuta, LLMaren neurona espezifikoak izango ditugu. Aktibazio oso txikia duten neuronak baztertzeko, hizkuntzaren batean aktibazio-probabilitate atalase batera iristen ez diren neuronak alde batera utziko ditugu. Beraz neurona bat k hizkuntzako espezifikoa izango da baldin eta LAPE baxuena duten neuronen $\%1$eko pertzentilean baldin badago eta ezarritako aktibazio-probabilitate atalasea gainditzen badu k hizkuntzan.

3 Esperimentuak

Atal honetan ereduen neurona espezifikoen banaketarekin lotutako emaitza esperimentalak aurkeztuko ditugu: erabili ditugun hizkuntzak, ereduetako neurona espezifikoen banaketak eta neuronen espezifikotasunaren analisia eta interpretazioa.

3.1 Esperimentuen diseinua

LAPE metrika kalkulatzeko, lehenengo esperimentaziorako hizkuntzen multzoa hautatu behar da. Gure esperimentuetan euskara, frantsesa, gaztelera eta ingelesa hautatu ditugu, ingelesa jatorrizko LLM fundazionalaren hizkuntza nagusia delako, gaztelera eta frantsesa baliabide askoko hizkuntzak eta gure ingurukoak direlako, eta euskara baliabide urriko eta gure hizkuntza gisa. Gainera, Llama-3.1-8B eredua baliabide urriko hizkuntzetara *continual pretraining*-aren bidez egokitzean neurona espezifikoen banaketa nola aldatzen den aztertu nahi dugu euskararen bitartez. Hizkuntza bakoitzeko milioi bat token (LLMak prozesatzen dituen datu-unitateak, hitzak edo hitzen zatiak direnak) dituen corpus bana osatu dugu Wikipediako artikuluekin esperimentuak egiteko.

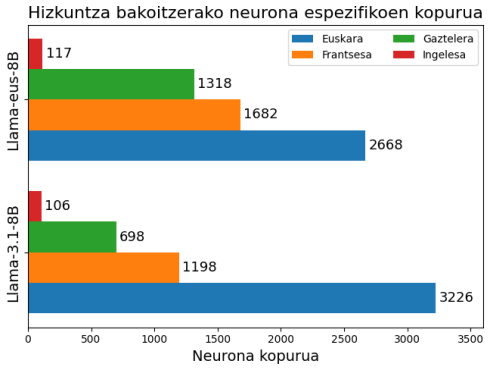
Ereduei dagokienez, Llama-3.1-8B eta Llama-eus-8B ereduek arkitektura berbera dute, izan ere, bigarrena lehenaren *continual pretraining* bidezko egokitzapena besterik ez baita. Arkitektura horretan 32 Transformer geruza daude, geruza bakoitzak bere FFN modulua du, eta FFN modulu bakoitzean 14336 neurona daude. Horrek esan nahi du denera $32 \cdot 14336 = 458752$ neurona daudela eredu bakoitzean.

Neurona espezifikoak LAPE metrikaren teknikarekin identifikatutakoan, interesgarria da aztertzea neurona horiek zenbateraino diren benetan hizkuntza horretako espezifikoak. Horretarako, bi gauza aztertu behar dira: neurona horien eragina hizkuntza espezifiko horretan bertan (benetan espezifikoak badira eragina handia izango da), eta baita beste hizkuntzetan duten eragina ere (txikia izango da egiazki espezifikoak baldin badira). Eragin hori neurtzeko ondorengo metodoa erabiliko dugu: hizkuntza bakoitzaren neurona espezifikoak desaktibatuko ditugu banan-banan (hau da, aktibazioa 0 izatera behartuko dugu), eta ondoren desaktibatze horrek hizkuntza bakoitzean duen portaera ikusiko dugu. Portaera neurtzeko aipatutako milioi bat tokeneko corpusetan kalkulatutako *perplexity* (PPL) puntuazioa erabiliko dugu, zeinak ereduak testu baten aurrean erakusten duen "harridura" neurtzen duen. Hau da, testu horrek probabilitate txikia baldin badu ereduarentzat, PPL handia izango du, eta alderantziz. Artikulu honetako esperimentuetan, hizkuntza bakoitzerako laginetan kalkulatuko dugu PPLa, hots, hizkuntza horretan testu zuzena sortzeko duen gaitasuna neurtuko dugu PPLaren bidez. Horrela, jakin dezakegu neurona espezifikoak desaktibatu ondoren zenbaiteraino ahultzen den LLMak hizkuntza horretan testu zuzena sortzeko duen gaitasuna.

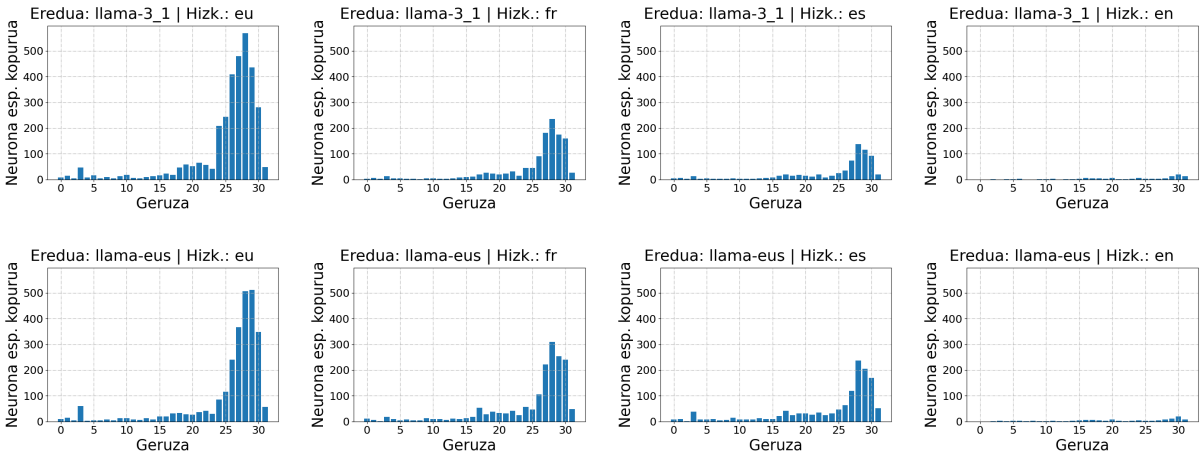
3.2 Neurona espezifikoen banaketa

LAPE metrika kalkulatu eta beharrezko kalkulu guztiak egin ondoren, eredu bakoitzak dituen neurona espezifikoen kopuruak 3 Irudian erakusten ditugu. Nabarmena da bi ereduetan ingeleserako dagoen neurona espezifikoen kopurua beste hizkuntzena baino askoz ere txikiagoa dela, eta euskarako, berriz, beste hizkuntzetarako baino neurona

3. Irudia: Llama-3.1-8B eta Llama-eus-8B ereduek hizkuntza bakoitzerako duten neurona espezifikoen kopurua.



4. Irudia: Neurona espezifikoen banaketa Llama-3.1-8B (goiko errenkada) eta Llama-eus-8B (beheko errenkada) LLMetan.



espezifiko gehiago daude. Oro har, joera ondoko hau dela esan daiteke: ereduak hizkuntza batean duen mailaren alderantziz proportzionala da hizkuntza horretarako duen neurona espezifikoen kopurua.

2 atalean aipatu bezala, Llama-eus-8B eredua Llama-3.1-8B ereduaren euskararako egokitzapena da, *continual pretraining* teknikaren bidez egokitua. Emaitzak ikusita, esan daiteke egokitzapen horretan euskararako espezifikoak diren neuronen kopurua jaitsi egin dela, eta beste hizkuntza guztiena igo. Hori bat dator aurretik aipatutako joerarekin: LLM bat hizkuntza batean hobeto funtziona dezan egokitzeak jaitsi egin baitu hizkuntza horretako neurona espezifikoen kopurua. Hala ere, esan beharra dago jaitsiera hori %17 ingurukoa dela, eta Llama-eus-8B ereduan euskarak jarraitzen duela neurona espezifiko kopuru handiena izaten.

Bestalde, neurona espezifikoak LLMen geruzetan nola banatzen diren dago irudikatuta 4 Irudian. Banaketaren itxurak antz handia du kasu guztietan: barrualdeko geruzetan neurona espezifiko gutxiago daude, 16. geruzatik aurrera igotzen hasten da kopuru hori, 29. geruza inguruan dago kopururik handiena eta azken 2-3 geruzetan berriro jaisten da.

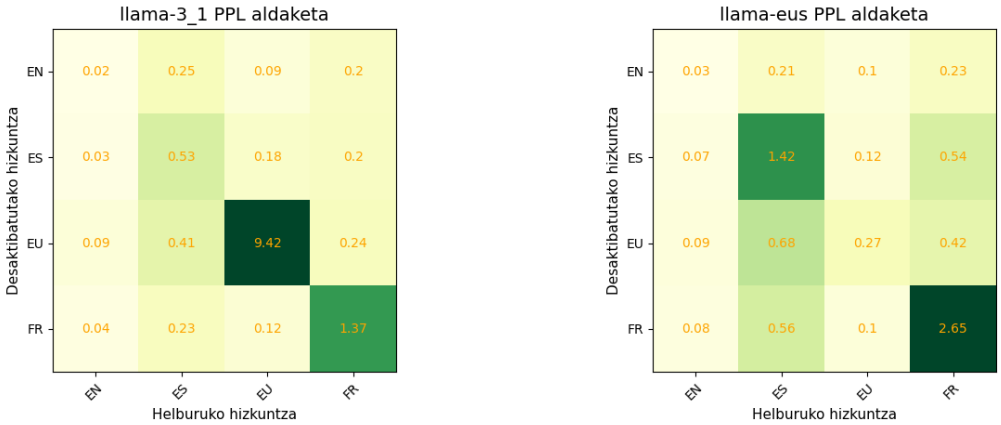
3.3 Neuronen espezifikotasuna

Neurona espezifikoak identifikatuta, ikus dezagun zenbateraino diren espezifikoak identifikatutako neuronak. Hizkuntza baten neuronak desaktibatuta ereduak duen portaeraren aldaketa neurtzeko, PPLaren aldaketa neurtuko dugu, era horretan zenbatetsiko dugu neurona horiek desaktibatuta hizkuntza bakoitza zenbateraino "ahaztu" zaion ereduari.

Identifikatutako neuronak benetan espezifikoak izango balira, desaktibatutakoan dagokien hizkuntzaren PPLa asko igoarazi beharko lukete, eta beste hizkuntza guztien PPLan ez luke eragin handirik izan beharko; hau da, 5

Irudiko diagonaleko zenbakiek handiak izan beharko lukete diagonaletik kanpo daudenen aldean. Ikusten dena ondorengo hau da: ingelesak bi ereduetan eta euskarak Llama-eus-8B ereduan ez dute portaera hori erakusten, hizkuntza horien neuronak desaktibatuta perplexitya ez baita gehien hizkuntza horretan jaisten. Llama-3.1-8B ereduan gazteleraren, euskararen eta frantsesaren kasuan eta Llama-eus-8B ereduan gaztelera eta frantsesaren kasuan neurona espezifikoak desaktibatzeak zentzuzko eragina du.

5. Irudia: Llama-3.1-8B eta Llama-eus-8B ereduetan hizkuntza bakoitzerako identifikatutako neurona espezifikoak desaktibatu ondoren hizkuntza bakoitzean neurtzen den PPLaren aldaketa.



4 Ondorioak

LAPE teknika erabilita, Llama-3.1 8B eta Llama-eus-8B ereduen euskara, frantses, gaztelera eta ingeles hizkuntzetarako FFN geruzetako neurona espezifikoak identifikatu ditugu. Neurona espezifikoen kopuru handiena kanpoaldeko geruzetan dagoela ikusi dugu, 29. geruzaren inguruan. Hipotesi hau proposatu izan da literaturan fenomeno hori azaltzeko: hasierako geruzetan eredua kontzeptuak "abstraktuan" pentsatzen hasten da, ez hizkuntza konkretu batean, eta geruzetan aurrera joan ahala kontzeptu hori hizkuntza jakin horretara ekartzen du (Tang *et al.*, 2024). Neurona espezifikoek kontzeptua hizkuntzara ekartzeko prozesu horretan hartzen dute parte, eta horregatik daude gehiago amaierako geruzetan. Gure emaitzek bat egiten dute hipotesi horrekin.

Bestalde, 3.3 azpiatalean ikusi dugu Llama-3.1-8B ereduan ingeleserako identifikatutako neurona espezifikoek ez dutela eragin handirik ereduak ingelesa ulertu eta sortzeko duen gaitasunean, eta gauza bera gertatzen da Llama-eus-8B ereduan ingeleserako eta euskararako neurona espezifikoekin. Dakigunez, Llama-3.1-8B eredua bereziki ingelesez erabiltzeko entrenatu da, eta Llama-eus-8B, berriz, euskaraz erabili ahal izateko egokitu da, baina ingelesez aurretik lortu dituen gaitasunak mantenduz. Beraz, era honetan interpreta daiteke aipatutako fenomenoa: eredua hizkuntza batean erabiltzeko bereziki egokitua baldin badago, hizkuntzari buruzko jakintza ereduko neuronen kopuru oso handi batera zabaltzen da eta neurona espezifikoek betetzen duten funtzioa (kontzeptu abstraktuak hizkuntza espezifikora ekartzearena) desagertu egiten da. Hipotesi horrek azalduko luke baita ere zergatik dauden beste hizkuntzetarako baino neurona espezifiko gutxiago ingeleserako bi ereduetan, eta zergatik jaisten den euskararako neurona espezifikoen kopurua Llama-3.1-8B eredua euskarara egokitzen denean.

5 Etorkizunerako planteatzen den norabidea

Aurrera begira, proposatutako hipotesiak baieztatzea gelditzen da, esperimentu gehiagorekin osatu behar baitira ikerketa honetan egindakoak. Ereduan nagusi diren hizkuntzen jakintza ereduko parametro gehiagotan gordetzen dela egiaztatu gabeko hipotesia da oraindik, eta arlo horretan egin daiteke ikerketa gehiago.

Azkenik, neurona espezifikoen kokapenari buruzko informazioa ereduak baliabide urriko hizkuntzetara egokitzeko tekniketan nola txertatu ikertzea gelditzen da, LLMen interpretazio mekanikoan egindako ikerketei ahalik eta probetxu handiena atera ahal izateko.

Erreferentziak

Chang, Hoyeon, Jinho Park, Seonghyeon Ye, Sohee Yang, Youngkyung Seo, Du-Seong Chang, eta Minjoon Seo. 2024. How do large language models acquire factual knowledge during pretraining? *arXiv preprint arXiv:2406.11813* .

Corral, Ander, Ixak Sarasua, eta Xabier Saralegi. 2024. Pipeline analysis for developing instruct llms in low-resource languages: A case study on basque. *arXiv preprint arXiv:2412.13922* .

Dai, Damai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, eta Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ed. by Smaranda Muresan, Preslav Nakov, eta Aline Villavicencio, 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Dubey, Abhimanyu, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, eta others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* .

Dubey, Shiv Ram, Satish Kumar Singh, eta Bidyut Baran Chaudhuri. 2022. Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing* 503.92–108.

Etxaniz, Julen, Oscar Sainz, Naiara Perez, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, eta Aitor Soroa. 2024. Latxa: An open language model and evaluation suite for basque. *arXiv preprint arXiv:2403.20266* .

Guo, Daya, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, eta others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* .

Lai, Wen, Viktor Hangya, eta Alexander Fraser. 2024. Style-specific neurons for steering llms in text style transfer. *arXiv preprint arXiv:2410.00593* .

Leng, Yongqi, eta Deyi Xiong. 2024. Towards understanding multi-task learning (generalization) of llms via detecting and exploring task-specific neurons. *arXiv preprint arXiv:2407.06488* .

Li, Yudong, Yuhao Feng, Wen Zhou, Zhe Zhao, Linlin Shen, Cheng Hou, eta Xianxu Hou. 2024. Dynamic data sampler for cross-language transfer learning in large language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 11291–11295. IEEE.

Mishra, Aditi, Utkarsh Soni, Anjana Arunkumar, Jinbin Huang, Bum Chul Kwon, eta Chris Bryan. 2023. Promptaid: Prompt exploration, perturbation, testing and iteration using visual analytics for large language models. *arXiv preprint arXiv:2304.01964* .

Nag, Arijit, Soumen Chakrabarti, Animesh Mukherjee, eta Niloy Ganguly. 2024. Efficient continual pre-training of llms for low-resource languages. *arXiv preprint arXiv:2412.10244* .

OpenAI, 2024. Learning to reason with llms.

Tang, Tianyi, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, eta Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. *arXiv preprint arXiv:2402.16438* .

Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* .

Yamaguchi, Atsuki, Aline Villavicencio, eta Nikolaos Aletras. 2024. How can we effectively expand the vocabulary of llms with 0.01 gb of target language text? *arXiv preprint arXiv:2406.11477* .

Zhang, Zhihao, Jun Zhao, Qi Zhang, Tao Gui, eta Xuanjing Huang. 2024. Unveiling linguistic regions in large language models. *arXiv preprint arXiv:2402.14700* .

6 Eskerrak eta oharrak

Eskerrik asko artikulu hau hobetzeko berrikuspenak eta iruzkinak egin dizkidaten Orai zentroko lankideei. Artikulu hau Bikaintek bekak lagundutako doktorego tesi bateko ikerketaren parte da.