



LLMetatik aplikazioetara: testua osatzetik aginduak jarraitzera

– Ander Corral

Informatikari Euskaldunen Bilkura, IEB 2024

2024/11/07

Motibazioa

Erronka: asistente bat zenbait ataza euskaraz egin ahal izateko.



LLMetatik aplikazioetara: testua osatzetik aginduak jarraitzera

Motibazioa

Gaur egun hori posible da ingelesez, gaztelaniaz, frantsesez...



... euskaraz puntu bateraino.

Open-source eta sistema komertzialen gaitasun linguistikoa euskaraz baxuagoa

Motibazioa



Provide a translation in Basque for this text:

La gran herramienta de la ciencia es su constante revisión, lo cual lleva al progreso científico.

> Zientziaren ~~tresna handia~~ tresnarik handiena bere etengabeko berrikuspena da, eta horrek aurrerapen zientifikoa dakar.



Zuzendu testu hau: Nik etorri naiz.

> **Testua zuzena da**, baina forma hobea emateko, "Nik etorri naiz." esateko, "Ni etorri naiz." erabil dezakezu. Biak dira zuzenak, baina bigarrenak egitura hobea du.

∞ Llama 3.1 70B

Idatz ezazu email bat nire lankideei. Gaia: IEB-n LLMak euskaratzeko egin dugun bidearen aurkezpena egingo dut

> Bai, hemen daukazu zure lankideei bidaliko **dizkien** mezua:

Gaiez,

Oso ongi,

Hurrengo astelehenean, IEBn egingo dut aurkezpen bat, eta **zuek ere bertan egon nahi dut**. LLMak euskaratzeko egin dugun bidearen aurkezpena egingo dut. Orain arte egin dugun lanaren emaitzak erakutsiko dizkizuet, eta aurrerantzean egin behar dugun lana **azaltuko** dizuet.

Ez dut galdu nahi, beraz, bertan egoten. Ordua eta tokiari buruz informazio gehiago igorriko dizuet hurrengo astelehenean.

Eskerrik asko zure parte-hartzeagatik eta laguntzagatik.

Agur,

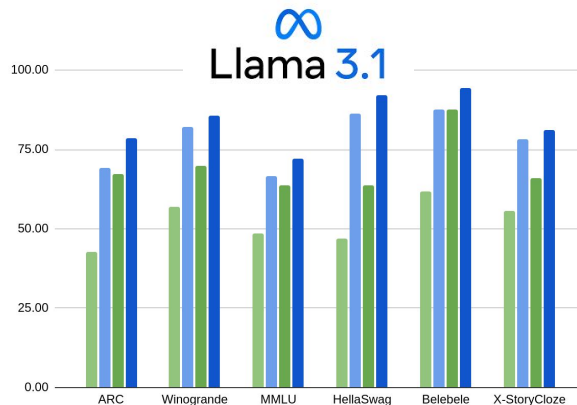
[Your Name]

LLMetatik aplikazioetara: testua osatzetik aginduak jarraitzera

Motibazioa

Bide-orria

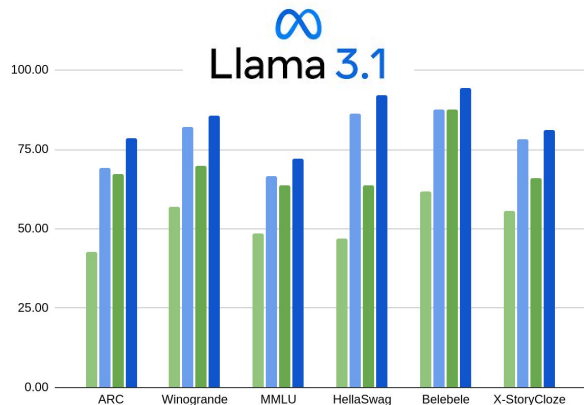
- 1) **LLM fundazional** on bat euskararako:
 - Gaitasun linguistikoak hobetzea.



Motibazioa

Bide-orria

- 1) **LLM fundazional** on bat euskararako:
 - Gaitasun linguistikoak hobetzea.
 - Ez oso astuna (<10B parametro).



~120 GB

8B

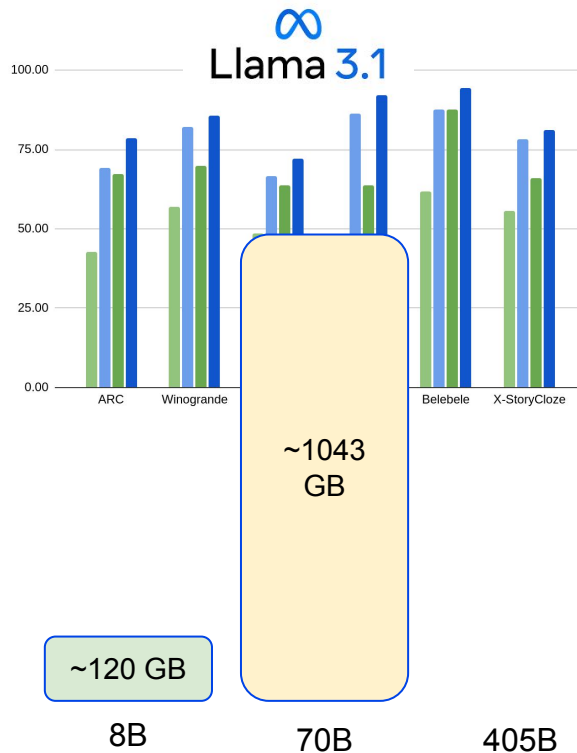
70B

405B

Motibazioa

Bide-orria

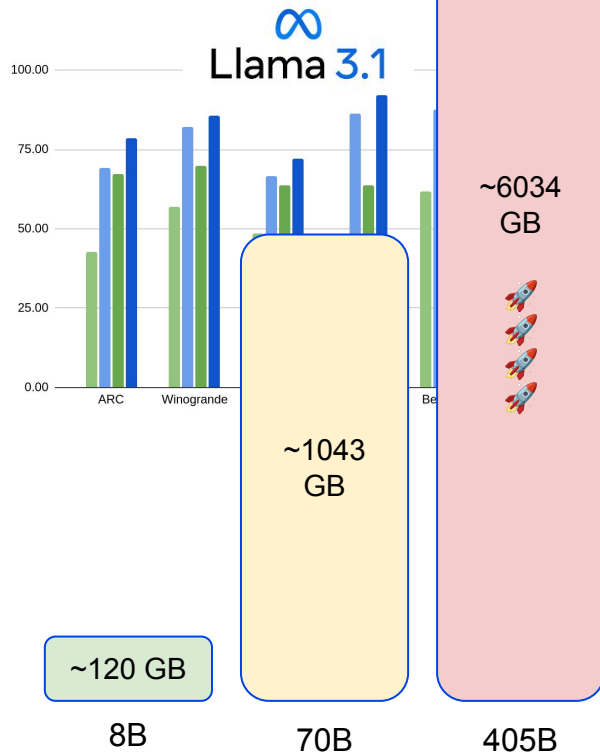
- 1) **LLM fundazional** on bat euskararako:
 - Gaitasun linguistikoak hobetzea
 - Ez oso astuna (<10B parametro).



Motibazioa

Bide-orria

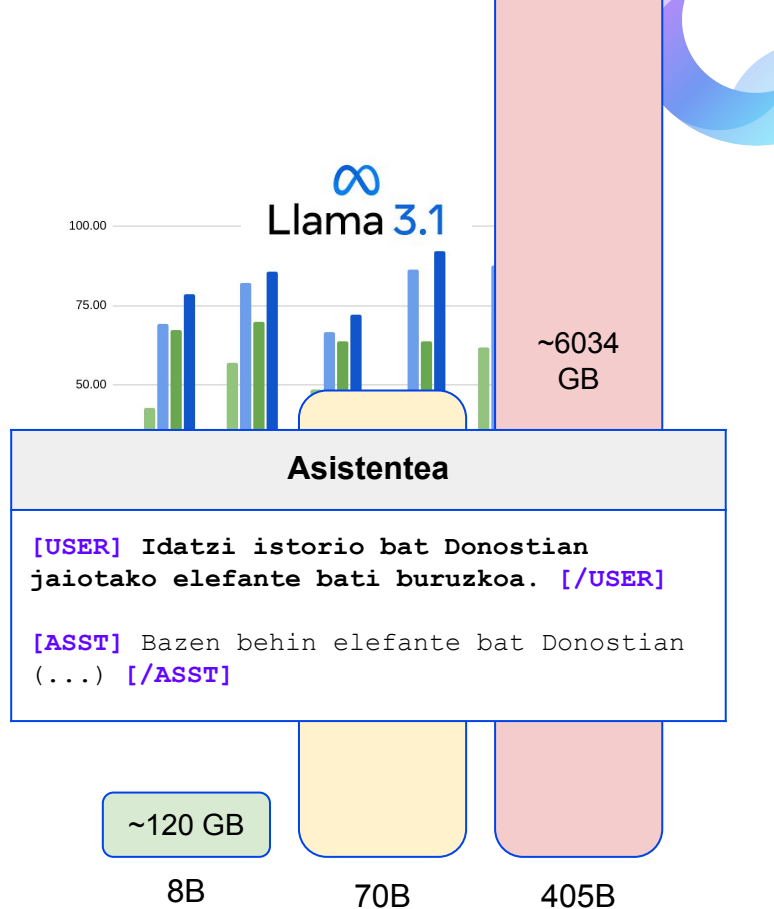
- 1) **LLM fundazional** on bat euskararako:
 - Gaitasun linguistikoak hobetzea.
 - Ez oso astuna (<10B parametro).



Motibazioa

Bide-orria

- 1) **LLM fundazional** on bat euskararako:
 - Gaitasun linguistikoak hobetzea.
 - Ez oso astuna (<10B parametro).
- 2) Euskararako **instruct LLMa** sortzea:
 - Asistente orokorra (hainbat ataza).
 - Espezializatua (ataza bakarra).



LLMetatik aplikazioetara: testua osatzetik aginduak jarraitzera

Motibazioa

Zergatik kode-irekiko LLMetan oinarritu?

Euskararako **instruct LLMak** sortzeko bide egokia:

- Egokitzapenak eta kontrola.
- Kostua.
- Pribatutasuna.
- Zenbait kasutan ereduak on-premise instalatu behar dira.

 Llama 3.1

Llama-eus-8B:

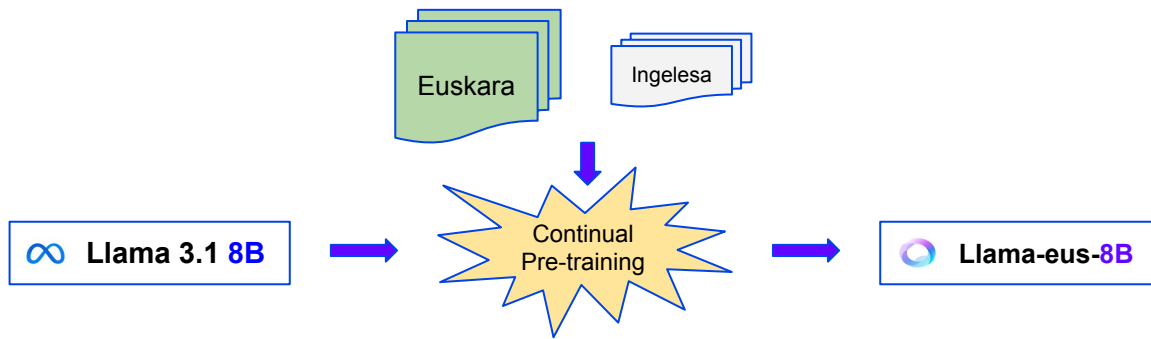
euskararako LLM
fundazionalaren
garapena

LLMei euskara irakasten

LLMei euskara irakasten

Erronka

LLMari euskararako gaitasun formala irakatea ingelesez ikasitako gaitasun funtzionala berrerabiliz.



Hizkuntza gaitasunak

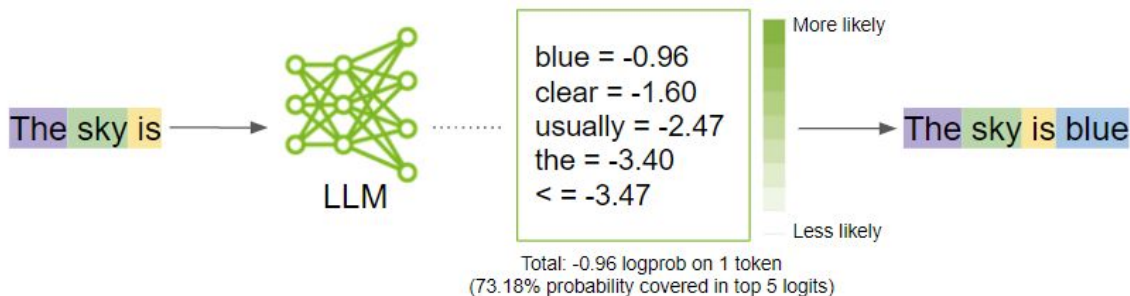
Formala: Gramatika eta hiztegien araberako hizkuntza ulertu eta sortzeko gaitasuna.

Funtzionala: Gaitasun formala beste era bateko gaitasunekin (arrazonamendua, munduaren ezagutza...) uztartuta hizkuntza egoera errealetan erabiltzeko gaitasuna.

Continual pre-training-a

Ereduak datu berriekin etengabe hobetu eta egokitzeko teknika bat.

- Ez da hasieratik entrenatzen (baliabide gutxiago behar dira).
- Aurreko datuetatik ikasitakoa berrerabili (transfer-learning).
- Modeloaren gaitasunak eguneratzen dira.



LLMei euskara irakasten

Entrenamendu-datuak



ZelaiHandi

Lizentzia libreko edukiak dituen eta kalitate handikoa den euskarazko corpusik handiena.

521 milioi hitz

⚠ *1.5B token Llama 3 tokenizer*



FineWeb

Ingelesezko CommonCrawl garbituta eta deduplikatuta.

15T token (500k doc subset)

⚠ *300M token Llama 3 tokenizer*

Euskararako LLM fundazionalaren garapena

Ebaluazioa

 Lau benchmark berri euskararako:

- **ARC_HT_eu_sample**
Lehen Hezkuntzako zientzia-galderak
- **Winogrande_HT_eu_sample**
Zentzuzko arrazoibidea
- **MMLU_HT_eu_sample**
Ezagutza orokorra: zientzia, humanitateak
- **HellaSwag_HT_eu_sample**
Esaldi-bukaeren desanbiguazioa

Estandarrak diren benchmark-en **250 adibideko laginen eskuzko itzulpena.**

Bestelakoak

BL2MP, **BasqueGLUE**, **Belebele**, **X-StoryCloze**,
EusProficiency, **EusReading**, **EusTrivia**, **EusExams**

Landareek eguzki-argia erabiltzen dute

A: lurzorua sortzeko.

B: mineralak sortzeko.

C: elikagaiak sortzeko.

D: ura sortzeko.

Zer ezaugarri dute zelula guztiek?

A: Energia behar dute.

B: Sexualki ugaltzen dira.

C: Eurek sortzen dute behar duten janaria.

D: Batetik bestera mugitzen dira.

m masako orga txiki bat geldirik d

Elastikoki talka egiten du 4m masa

abiadurako orga handi batekin. Orga handiak

bere energia zinetikoaren erdia galtzen du. Orga

txikiak abiadura hau du orain:

A: 1,41v

B: v

C: 2v

D: 4v

Emakumeak bikini beretzat hartu zuen, baina elastikoa itzuli egin zuen, [MASK] tamaina egokia zuelako.

A: bikiniak

B: elastikoak

Euskararako LLM fundazionalaren garapena

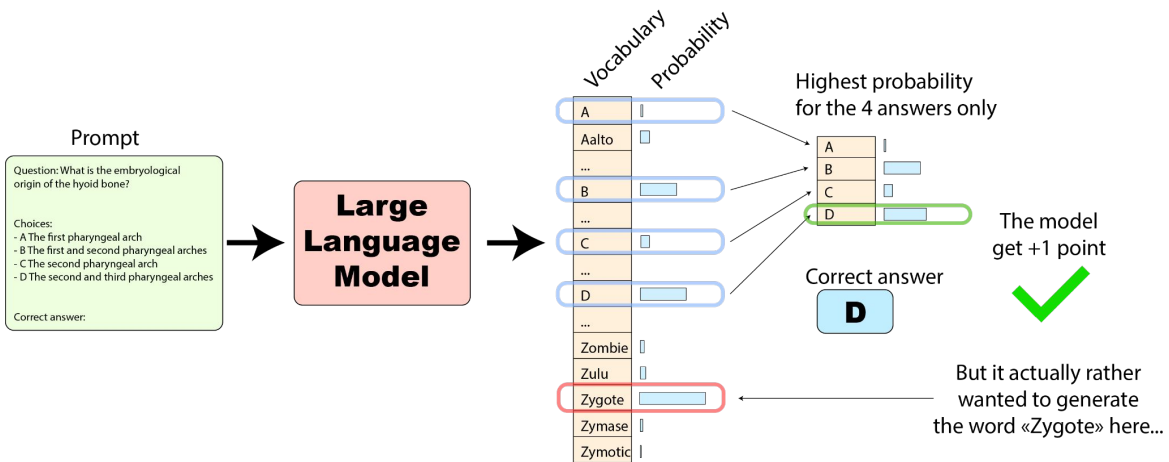
Ebaluazioa

LM Evaluation Harness erramintarekin

Automatikoa

Test motako galderak

In-context learning



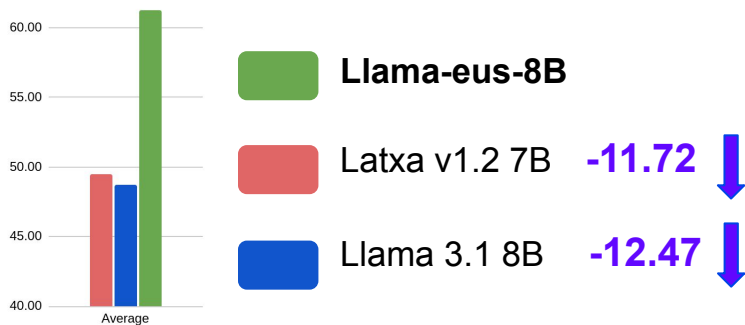
Kasu! LLMen potentziala neurtzeko, ez du zertan LLMak bueltatuko lukeena izan

Euskararako LLM fundazionalaren garapena

Emaitzak

10B parametro baino gutxiagoko LLMak:

- Llama-eus-8B modelorik hoberena.

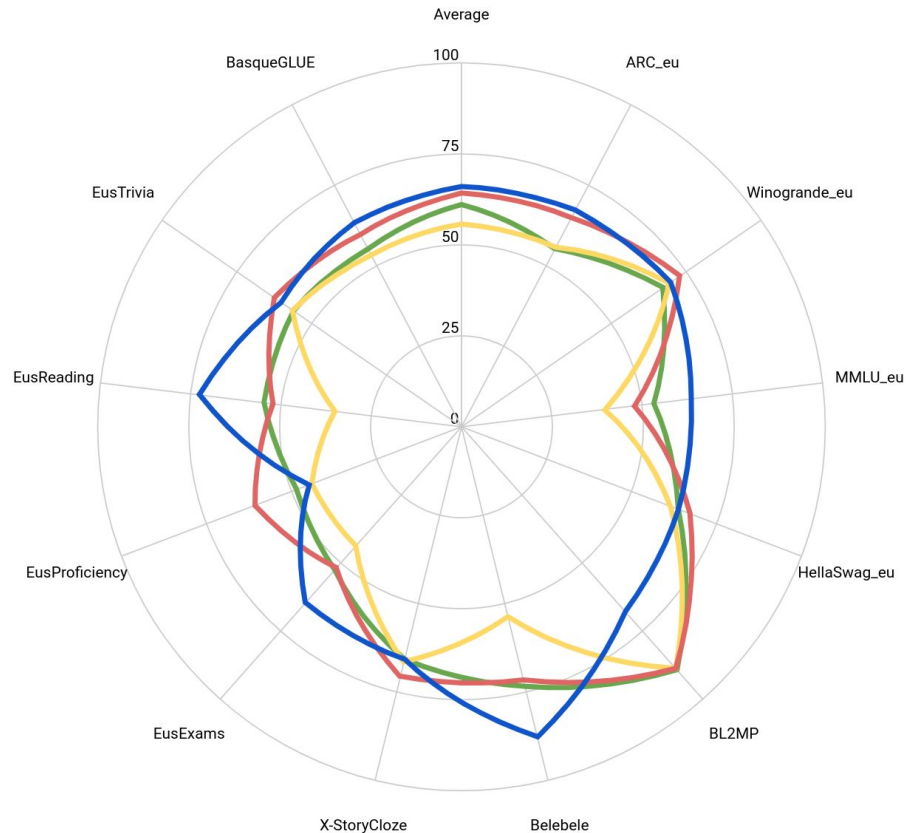
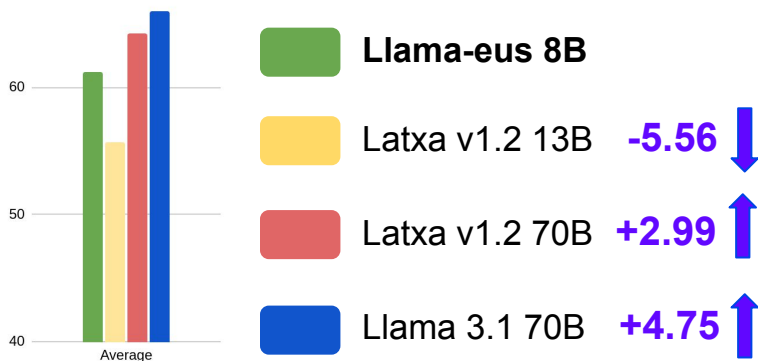


Euskararako LLM fundazionalaren garapena

Emaitzak

10B parametro baino gehiagoko LLMak:

- Latxa v1.2 13B gaitzen da (x2 param).
- Konpetitiboa 70B LLMekiko (x10 param).

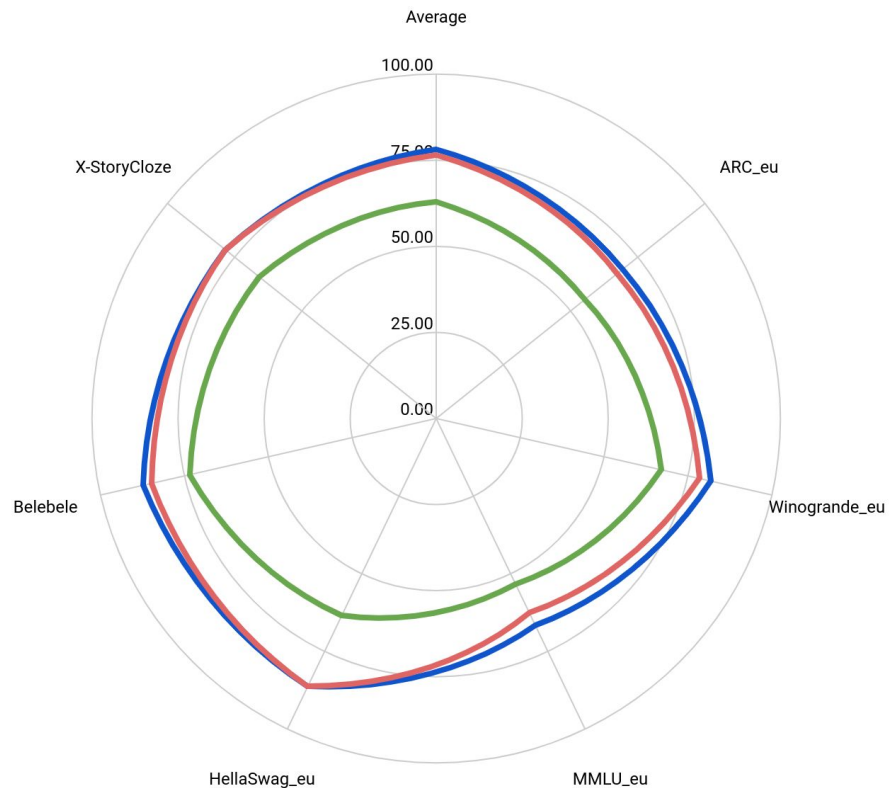
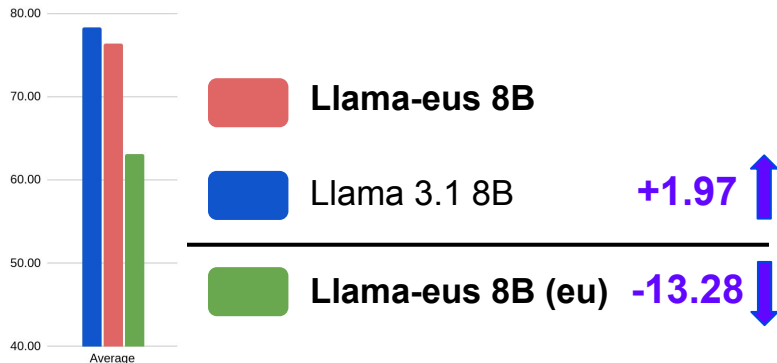


Euskararako LLM fundazionalaren garapena

Emaitzak

Eta ingelesez zer?

- Degradazio gutxi.
- Alde nabarmena euskararekiko.



Euskararako asistentearen garapena

Euskararako asistentearen garapena

Euskararako asistentea

Erronka: euskararako helburu orokorreko asistentea sortzea.

Ataza asko:



Euskararako asistentearen garapena

Instrukzio fine-tuning-a

Zer da?

LLM fundazional bati erabiltzailearen aginduak jarraitzen irakastea.

“Testua osatzetik aginduak jarraitzera”

Nola?

Instrukzio-erantzun bikoteekin egokitzen da.

LLM fundazionala

Zein da Gipuzkoako hiriburua?

> *Hiriburu izatearen ideia filosofikoki aztertuta, hiri bat bere eraikin, kale eta azpiegiturak baino gehiago da. Hiri batek, bereziki hiriburuak, ikusezina den zerbait biltzen du: sinbolismoa. Donostia, sinbolikoki, euskal erresistentziaren eta herri izaeraren irudi bilakatzen da askorentzat. [...]*

Instrukzio
fine-tuning-a

Asistentea

Zein da Gipuzkoako hiriburua?

> *Gipuzkoako hiriburua Donostia da.*

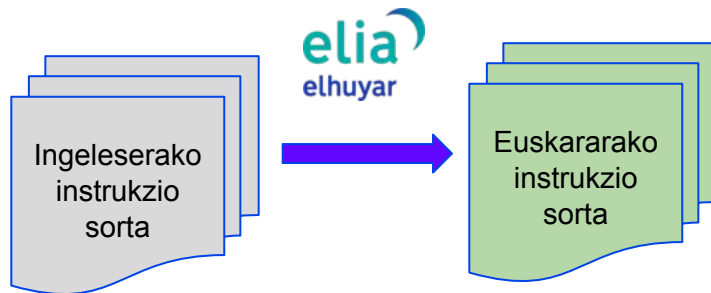
Euskararako asistentearen garapena

Instrukzio-sortak euskaratzen

Arazoak:

- Euskararako ez dago kalitatezko eta librea den datu-multzorik.
- Euskararako instrukzio sortak eskuz sortzea oso garestia da.

Soluzioa: itzulpen automatikoa baliatuz, euskararako instrukzio sorta handia eta anitza sortzea.



Euskararako asistentearen garapena

Instrukzio-sortak euskaratzen



NoRobots

Eskuz sortutako kalitate handiko
instrukzio dataset-a

10k instrukzio



SlimOrca

GPT4 bidez automatikoki sortutako
eta filtratutako instrukzio dataset-a

500k instrukzio

⚠️ Automatikoki sortuta

Euskararako asistentearen garapena

Instrukzio-sortak euskaratzen

Mezu elektronikoa bat idatzi astronauta batek espaziotik NASARA mezu bat bidalko balu bezala. Demagun astronautek posta elektronikoa dutela espazioan. Mamuak aurkitu dituzte ilargian. Tonua kezkatua eta apur bat zurbila izan behar da.

> NASA agurgarria,

Mesedez, irakurri mezu hau, egoerak berehalako arreta eskatzen baitu.

Ilargiaren ohiko misioan, gure irudimenik basatiena baino harantzago doan fenomeno aurkitu dugu. Sinestezina badirudi ere, ilargian mamuak daudela iradokitzen duela ebidentzia aurkitu dugu. Bai, zuzen irakurtzen duzu: mamuak!

[...]

Zuen erantzunaren zain,

Astronauta

Chicagoko eta Milwaukeeko kirol talde profesional nagusien zerrenda egin dezakezu?

> -Chicagoko Bears

-Chicago Bulls

-Chicago White Sox

-Milwaukee Brewers

-Milwaukee Bucks

-Milwaukee Wave

Idatzi istorio bat Donostian jaiotako elefante bati buruzkoa.

> Bazen behin elefante bat Donostian [...]

Zein da Taj Mahalaren esanahia eta nork agindu zuen hura eraikitzea?

> Indiako Agran dagoen Taj Mahal munduko egitura ospetsu eta arkitektonikoki esanguratsuenetakoa da. 1632an, Shah Jahan enperadore mughalak Mumtaz Mahal emazte kutunaren mausoleoa egiteko mandatua eman zion; emaztea erditzean hil zen. Taj Mahal marmol zuriko eta harri erdibitxiko arkitektura-miraria da, estilo arkitektoniko pertsiarren, islamiarren eta indiarren eraginak islatzen dituena. 1983an UNESCOren munduko ondare izendatu zuten, eta Indiaren historia aberatsaren ikur izaten jarraitzen du. [...]

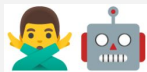


Euskararako asistentearen garapena

Ebaluazioa

Arazoa: sorkuntza-atazak ebaluatzeko ez dago metodologia automatiko eraginkorrik.

Metodologia: Eskuzko ebaluazioa egin da, fidagarriena xehetasunak harrapatzeko.



NoRobots test

100 instrukzioko ausazko lagin bat

Euskarara eskuz itzuli da

10 instrukzio-mota

Erantzunak **hiru kategoriatan** sailkatu dira eskuz:

Zuzena	Erantzuna egokia eta zuzena
Partzialki zuzena	Erantzuna ontzat eman daiteke baina informazio edo zehaztasun gabeziak ditu
Okerra	Erantzunak ez ditu instrukzioaren eskakizunak betetzen

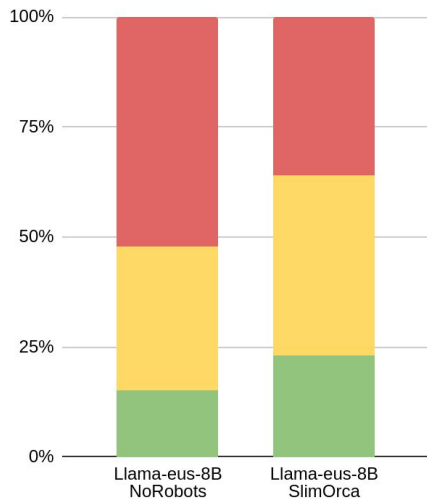
Euskararako asistentearen garapena

Emaitzak

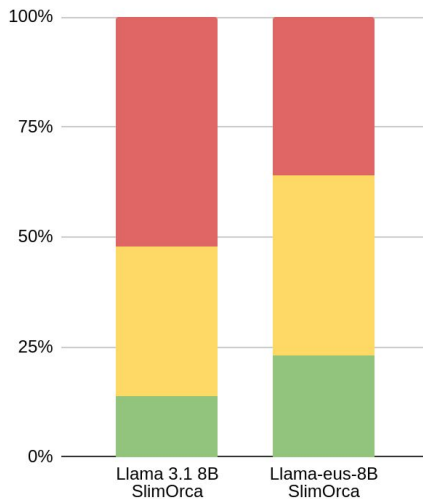


Llama-eus-8B instruct

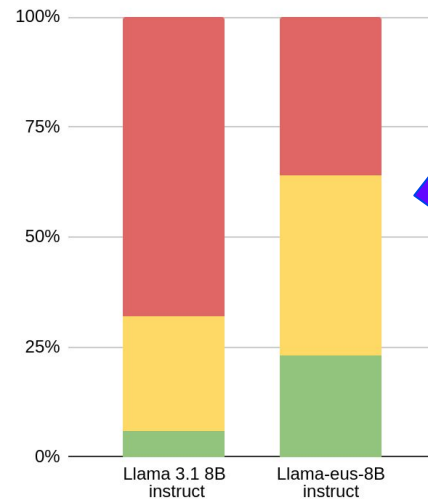
Datu-multzoaren aukeraketa



Euskara aurre-entrenamenduan



vs. Llama 3.1 8B instruct



Euskararako asistentearen garapena

Reinforcement learning-a

Zer da?


Instruct LLM bat fintzea **erabiltzaileen feedback**-arekin.

Prozesu iteratiboa da normalean.

Nola?


Erantzun zuzen eta okerrekin egokitzea.

Ultrafeedback datu-multzoa automatikoki itzuli da.

Asistentea 

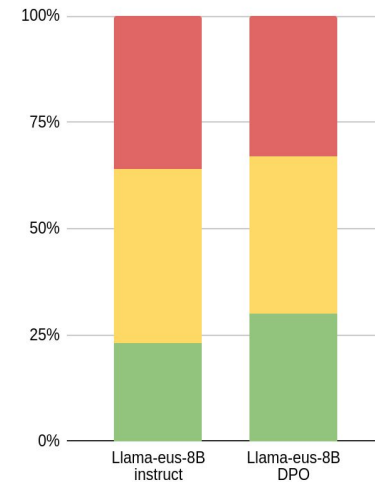
Zein da Gipuzkoako hiriburua?

> *Gipuzkoako hiriburua Donostia da. Donostia da Gipuzkoako hiririk handiena 187.000 biztanlerekin.*

Asistentea 

Zein da Gipuzkoako hiriburua?

> *Gipuzkoako hiriburua Donostia da.*



Itzulpen automatikoa:

instrukzio fine-tuning
espezializatua

Itzulpen automatikoa: Instrukzio fine-tuning espezializatua

Itzulpen automatikoa

Helburua: Llama-eus-8B euskararen sorkuntza behar duen ataza batera egokitzea.

Euskara helburu duten itzulpen-norabideak:

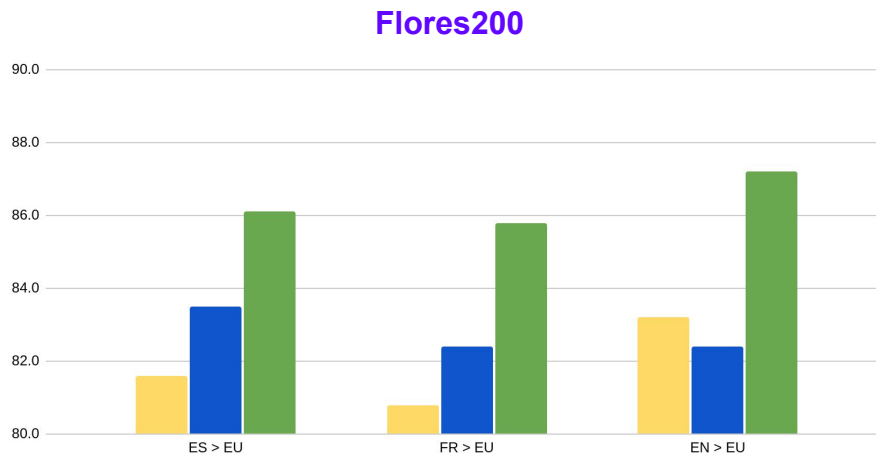
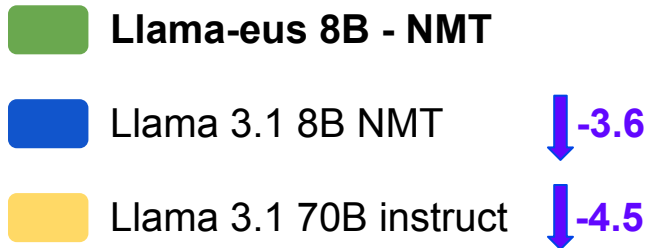
- **Gaztelania > Euskara**
- **Frantsesa > Euskara**
- **Ingelesa > Euskara**



Itzulpen automatikoa: Instrukzio fine-tuning espezializatua

Emaitzak

- LLMari euskara irakastea ezinbestekoa da (Llama-eus > Llama 3.1).
- Llama 3.1 70B baino hobea (x10 param).

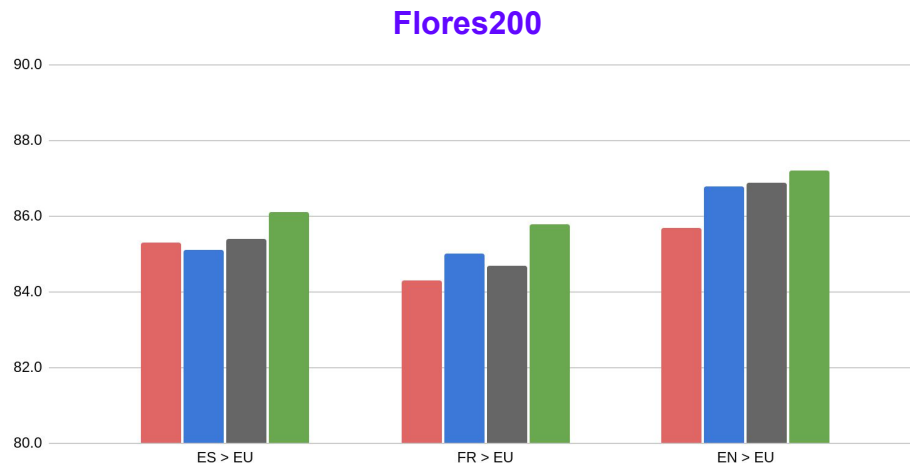
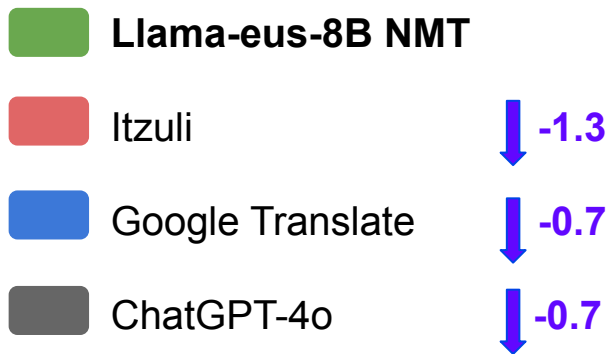


* Sistemen itzulpen-kalitatea COMET22 metrikarekin neurtu da

Itzulpen automatikoa: Instrukzio fine-tuning espezializatua

Emaizak

Sistema komertzialekin alderatuta?



* Sistemen itzulpen-kalitatea COMET22 metrikarekin neurtu da

Emaitzak

Llama-eus-8B NMT

Depuis les résultats du second tour des élections législatives anticipées, le 7 juillet dernier, les deux têtes d'affiche du Rassemblement national ont l'intime conviction que la situation politique est intenable.

> Uztailaren 7ko hauteskunde legegileetako bigarren itzulpen emaitzak ikusita, Rassemblement national-eko bi buruzagia uste dute egoera politikoa ezin dela mantendu.

Llama-eus-8B NMT

¿Son las historias de ChatGPT tan buenas como las de un humano? Un nuevo estudio lo analiza. Un estudio reciente analiza las sorprendentes diferencias entre las historias generadas por inteligencia artificial y las creadas por humanos, y lo que esto significa para el futuro de la IA.

> ChatGPTren istorioak gizakienak bezain onak dira? Ikerketa berri batek aztertu du hori. Ikerketa berri batek aztertu ditu IAk sortutako istorioen eta gizakiek sortutako arteko alde harrigarriak, eta horrek narraziogintzaren etorkizunerako zer esan nahi duen.

Llama-eus-8B NMT

Nearly all of US states are facing droughts, an unprecedented number. More than 150 million people and 318m acres of crops are affected by droughts after summer of record heat.

> AEBko ia estatu guztietan lehortea dira nagusi, inoiz ez bezalako kopuruan. 150 milioi pertsona baino gehiago eta 318 milioi hektarea baino gehiagoko uzta lehorteen eraginpean daude, udako bero-errekorraren ondoren.

Ondorioak

LLMetatik aplikazioetara: testua osatzetik aginduak jarraitzera

Ondorioak

Euskararako LLM fundazionala:

- **Kode irekiko** LLM fundazionalak egokitzea bide **eraginkorra** euskararako LLMak sortzeko.
- LLM fundazional on bat euskararako: **Llama-eus-8B**
 - **Gaitasun linguistikoak** nabarmen hobetu dira: **+ 12.5 puntu**
 - 521M hitzeko **euskarazko corpusa** erabilita.
 - **Arina** (<10B parametro)
- Ingelesaren aldean diferentzia esanguratsua: **- 13.28 puntu**
 - **LLM fundazionala hobetzeko beharra**

LLMetatik aplikazioetara: testua osatzetik aginduak jarraitzera

Ondorioak

Instruct LLMekin lehen esperimentazioa:

- **Asistente orokorra: Llama-eus-8B instruct**
 - Hobekuntza esanguratsua (+24 puntu)
 - Ingelesetik automatikoki itzultitako entrenamendu instrukzioak erabilia
 - Oraindik nabarmen hobetzeko beharra.
- **Espezializatua (NMT): Llama-eus-8B NMT**
 - Euskarara itzultzen emaitzarik onenak.
- Datuen falta (corpusak, instrukzio-sortak...) desabantaila nabarmena hizkuntza handien aldean.

“Testua osatzetik aginduak jarraitzera”



Galderarik?

Baliabideak



Llama-eus 8B



ZelaiHandi