Euskara hizkuntza-eredu handietan eta txatbotetan bidea eginez

Badira jada bi urte edozein galdera erantzuteko eta mota guztietako atazak egiteko gai diren txatbotak agertu zirela, eta geroztik ChatGPT, Gemini, Copilot, Claude eta enparauak toki guztietan daude eta oso erabiliak dira. Tresna horien oinarrian hizkuntza-eredu handiak edo LLMak (Large Language Models) daude. Euskara eredu horietan egiten ari den bidea azalduko dugu artikulu honetan. Eta txatboten eta LLMen funtzionamendua ere azalduko dugu, euskarak horietan aurrera egiteko dauden zailtasunak ulertu ahal izateko batetik, baina baita ere garrantzitsua delako nonahikoa eta alor askotan gero eta beharrezkoagoa den edozein teknologiaren funtzionamendua ulertzea.

Duela ia bi urte *Adimen artifizial sortzailearen booma* izenburuko artikulua idatzi genuen txoko honetan. Eta ezin esan dezakegu ordutik hona leherketa haren eragina apaldu denik; alderantziz, boom haren hedatze-uhina etengabe eta noranahi zabaldu da. Orduan aipatutako Adimen Artifizial edo AA sortzaileko sistemak testua edo irudia sortzeko gai ziren; bada, geroztik ikusi ditugu musika sortzen duten sistemak (Suno, adibidez), bai eta bideoa ere (OpenAI-ren Sora, adibidez). Txatbot berri mordoa agertu dira: Microsoft-en Copilot, Anthropic-en Claude, Google-ren Gemini, Meta-ren Meta AI, Perplexity, Jasper AI... Eta orobat ugaritu dira LLM (Large Language Model) edo hizkuntza-eredu handi berriak edo lehengoen bertsio berriak: PaLM, GPT-4, Grok, Gemini, Claude... Halaber, lizentzia libreko LLMetan ere asko aldatu da panorama, OpenAssistant, Mixtral, Gemma, Qwen eta, batez ere, Metaren LLaMaren agerpenarekin.

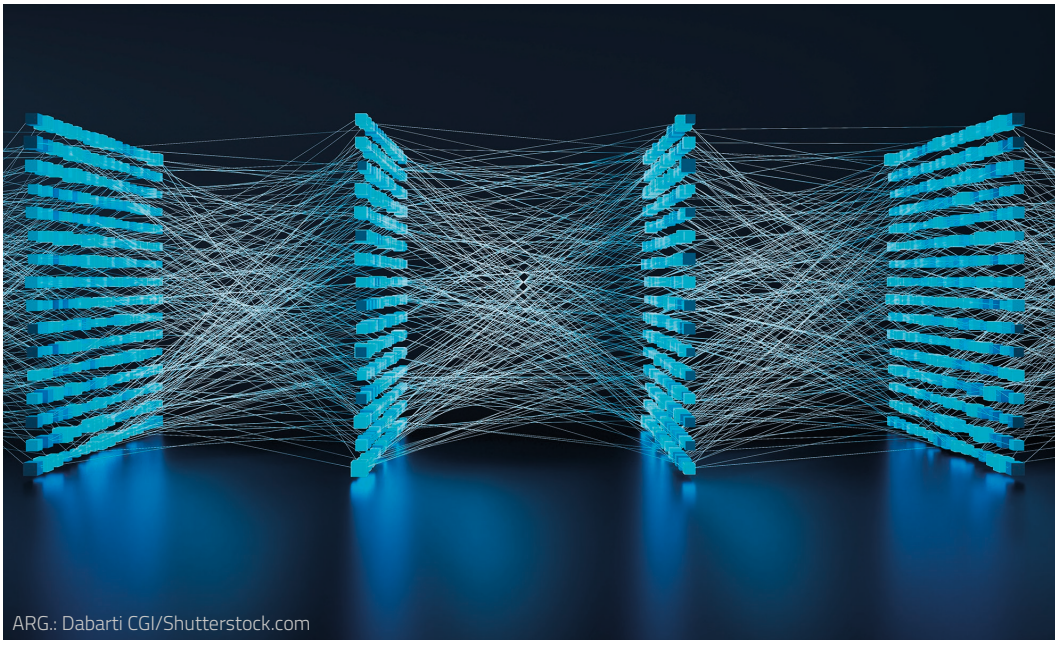
LLM edo Hizkuntza Eredu Handiak

Ez da erraza izen- eta kontzeptu-zopa horren hariari jarraitzea. Baina gutxienez komenigarria eta interesgarria ere bada jakitea zer den LLM bat, zer txatbot bat, eta nola garatzen diren.

Esan daiteke duela bizpahiru urtetik hona aro berri bat bizitzen ari garela hizkuntza- eta hizketateknologietan, hizkuntza-eredu handien edo LLMen aroa hain zuzen ere. LLMak sare neuronal sakon mota bat dira, gai direnak, batetik, orain arte ebatzi gabeko problema askori ongi erantzuteko (testuen sorkuntza automatikoa, edozein motako galderei erantzutea, programa informatikoak idaztea...), eta, bestetik, beste sare neuronal sakon mota batzuekin jada egiten ziren ataza asko haiek bezain ondo edo hobeto egiteko (itzulpen automatikoa, laburpen automatikoa...). Finean, testuzko edozein lan LLMen bidez egin daiteke gaur egun (eta, kasu askotan, egiten da).

LLMak sare neuronal sakonak dira, *transformer* motakoak, eta horren barruan gehienak *decoder* soil klasekoak izan ohi dira. Baina egituraz harago, beste hainbat ezaugarri garrantzitsu dituzte. Batetik, sare erraldoiak dira: sarrerako nodo asko dituzte, eta tarteko geruza asko bakoitza nodo askorekin; nodo horien arteko loturak edo sareak ikasi behar dituen parametroak milaka milioi izan ohi dira. Bestetik, testu hutsekin entrenatzen dira, baina testu-kopuru ikaragarri handiekin, milioika milioi hitz arterainoko

Igor Leturia Azkarate
Informatikaria eta ikertzailea



ARG.: Dabarti CGI/Shutterstock.com

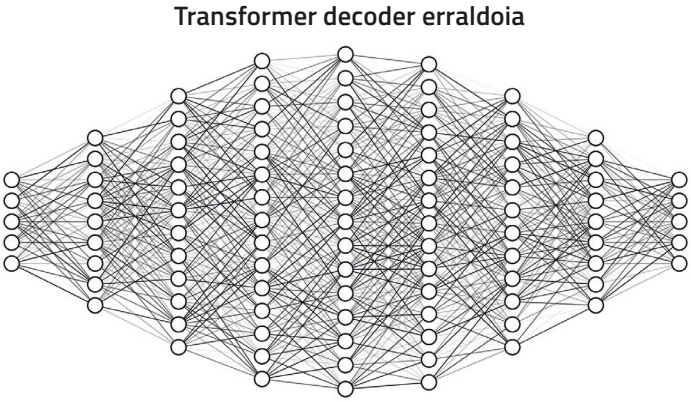
testu-bildumekin. Azkenik, eleaniztunak izan ohi dira, hau da, hizkuntza askotako testuekin entrenatzen dira eta hizkuntza horietan guztietan funtzionatzen dute (baina hobeto batzuetan beste batzuetan baino, geroago ikusiko den bezala).

LLMen egitekoa, printzipioz, gauza bakar eta oso sinple bat da: hitz-segida bat emanik, entrenamenduan ikusi dituen hitz-sekuentziak kontuan izanda, hurrengo hitz probableena aurresatea. Hau da, sarreran "urrutiko intxaurrak hamalau," segida ematen badiogu, berak "gerturatu" itzuli beharko luke. Eta hori besterik ez du egiten LLM batek. Kontua da jarraian "urrutiko intxaurrak hamalau, gerturatu" ematen badiogu, "eta" itzuliko duela, berriz eginez gero "lau"... Eta horrela, auto-erregresio deritzon prozesu horren bidez, testu edo erantzun luzeak sortzen jar ditzakegu. Gainera, esan bezala, sarreran nodo ugari izan ditzakete, are milaka batzuk ere, eta horrek ahalbidetzen du sarrerako testuak edo eskaerak ("prompt"-ak, arloko terminoa baliatuta) oso luzeak eta konplexuak izatea.

Egiatan, LLMek ez dute hitzekin funtzionatzen, zenbakiekin baizik. Eta hizkuntza askotako hitz guztiak adierazteko zenbaki gehiegi behar direnez, "token" deritzen hitz-zatiekin egiten dute lan. Baina azalpenak errazteko, hitzak hartzen dituztela esango dugu.

Beraz, laburbilduz, hau da LLM bat: transformer motako sare neuronal erraldoi bat (normalean *decoder* klasekoa), testu luze bat emanda hurrengo hitza aurreikusten duena, eta horretarako hizkuntza askotako testu askorekin aurre-entrenatu dena (aurrerago ikusiko dugu zergatik "aurre"-entrenamendu). Oinarrizko egoera horretan dauden LLMek beste hainbat izen ere jasotzen dituzte: eredu fundazional, GPT (*Generative Pre-Trained Transformer*), hizkuntza eredu autoerregresibo...

1. FASEA. AURRE-ENTRENAMENDUA

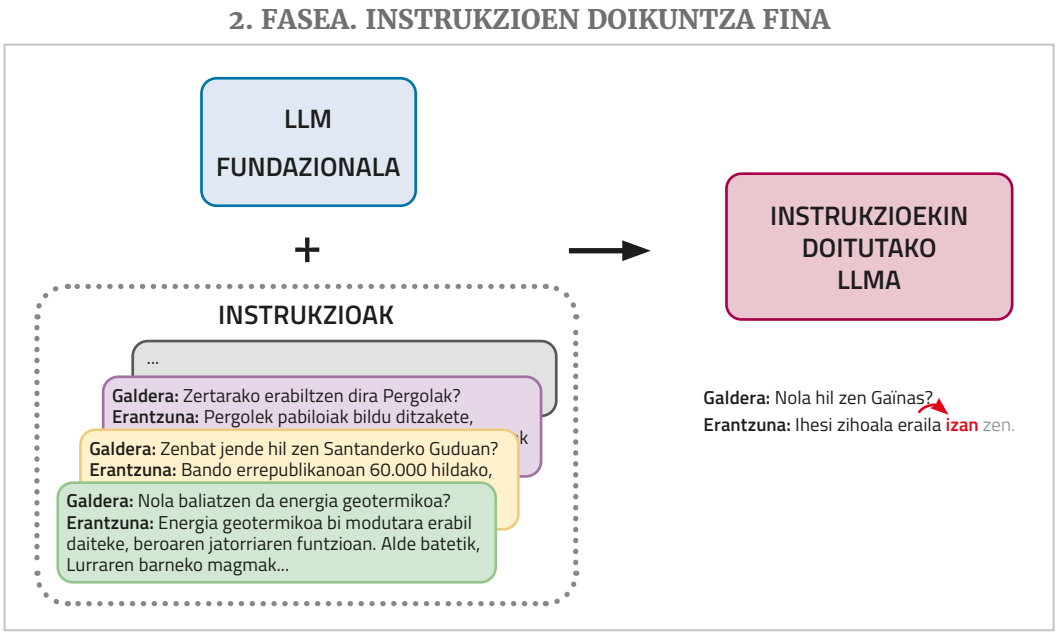


Gaïnas Teodosio I.a Handiaren eta honen seme eta oinordeko izan zen Arkadioren erregealdietan zerbitzatu zuen Bizantziar Inperioko buruzagi godo anbiziotsu bat izan zen. Gaïnasek, bere karrera militarra oinezko soldadu bezala hasi zuen eta, aurrerago, Teodosioren armadako barbaroen talde bat zuzendu zuen Flavio Eugenio usurpatzailearen aurka, 394an. 395ean Flavio Estilikon eta Eutropiorekin indarrak bateratu zituen Ekialdeko pretorioaren prefektua zen Rufino erorarazteko. 399an, Leoren ordezko izan zen magister militum karguan, azken honek Tribigildo buruzagi ostrogodoak zuzendutako inbasioak gelditzeko saiakeran huts egin ondoren.

Sarearen eta entrenamendu-bildumaren tamaina handiak direla-eta, sortzen dituen testuak linguistikoki zuzenak izan ohi dira, morfologia, sintaxi eta semantika mailan egokiak, eta munduari buruzko jakintza handia eta beste zenbait gaitasun erakusten ditu orokorrean.

LLMetatik txatbotetara

ChatGPT edo Gemini bezalako txatbot batek horrelako LLM bat du oinarrian, baina oraindik beste bi entrenamendu-urrats behar ditu: instrukzioen doikuntza fina eta gizakion hobespenekin lerrokatzea. LLM batek, gorago aipatu bezala, jarraipena ematen dio sarreran ematen zaion testu bati. Baina txatbotei galderak edo ataza bat egiteko eskaerak ematen zaizkie. Horrelakoen aurrean, baliteke LLMak ongi erantzutea entrenamenduko testuetan antzeko galdera-erantzun edo eskaera-soluzioak ikusi baditu. Baina entrenamenduko corpusetan ez da horrelako asko egoten. Horregatik, txatbot-lana hobeto betetzeko, LLMak instrukzioen doikuntza fina (*instruction fine-tuning*, ingelesez) deritzon beste entrenamendu-fase bat behar du. Entrenamendu horretarako, txatbotak egitea nahi den ataza-mota ez-



berdinetarako instrukzioen bilduma bat osatu behar da, hau da, eskaera-soluzio pareen adibideak: galderak erantzunekin, laburpen-eskaerak laburpenekin, testu okerrak zuzentzeko eskaerak dagozkien zuzenketekin, itzulpen-eskaerak itzulpenekin, testuak sortzeko eskaerak testuekin... Horrela, eskaera-mota horiei erantzun egokia ematen ikasten du LLMak.

Bukatzeko, hobeto funtzionatu dezan, gizakion hobespenekin lerrokatzea (*alignment with human preferences*, ingelesez) deritzon entrenamendu-fase bat behar da. LLMari hainbat ataza egin ditzan eskatzen zaio, bakoitzarentzat erantzun posible bat baino gehiago eskatuz, eta gero, hainbat pertsonak erantzun horiek ebaluatzen dituzte. Puntuatutako erantzun horiek baliatzen dira azken entrenamendu horretan, eta horren bidez lortzen da txatboten erantzunak hobeto bat egitea ("lerrokatzea", arloko terminoa baliatuta) gizakion nahiekin edo logikarekin. Urrats hori egiteko, teknika ezberdinak daude, hala nola giza berrelikadura bidezko errefortzu ikasketa (ingelesez, *Reinforcement Learning from Human Feedback* edo RLHF) edo zuzeneko lehentasunen optimizazioa (ingelesez, *Direct Preference Optimization* edo DPO).

Euskarazko LLM propioak eraikitzen

LLMak edo txatbotak eleaniztunak direla esan dugu gorago; izan ere, orokorrean hainbat hizkuntzatako testuekin entrenatzen dira. Baina ez dute berdin funtzionatzen hizkuntza guztientzat, entrenamenduan ikusitako testu-kopurua ez baita berdina: testu gehien-gehienak ingelesez dira, beste hizkuntzetakoak (baita beste hizkuntza nagusietakoak ere) askoz gutxiago dira, eta zer esanik ez hizkuntza txikietakoak. Hala ere, alde horiek egonik ere, hizkuntza guztiak nahiko ongi ikastera irits daiteke LLM bat, transferentzia bidezko ikasketa (*transfer learning*, ingelesez) deritzon propietatea dela eta. Frogatuta dago sare neuronalek badaukatela propietate hori, eta, horren bidez, nolabait domeinu edo hizkuntza batean edo batzuetan eskuratutako jakintza baliatzen dute beste domeinu edo hizkuntza bat datu gutxiagorekin ikasi ahal izateko; hizkuntzen kasuan, gainera, datu gutxiago beharko dira familia bereko hizkuntza bat jada menperatzen badu (gizakiokin ere antzera gertatzen da neurri batean). Hala, aurreko artikulu hartan ChatGPT-k euskaraz nahiko ongi egiten zuela baina oraindik hobetzeko asko zuela bagenioen ere, GPT 3.5etik GPT 4ra pasatzean, hobekuntza nabarmena

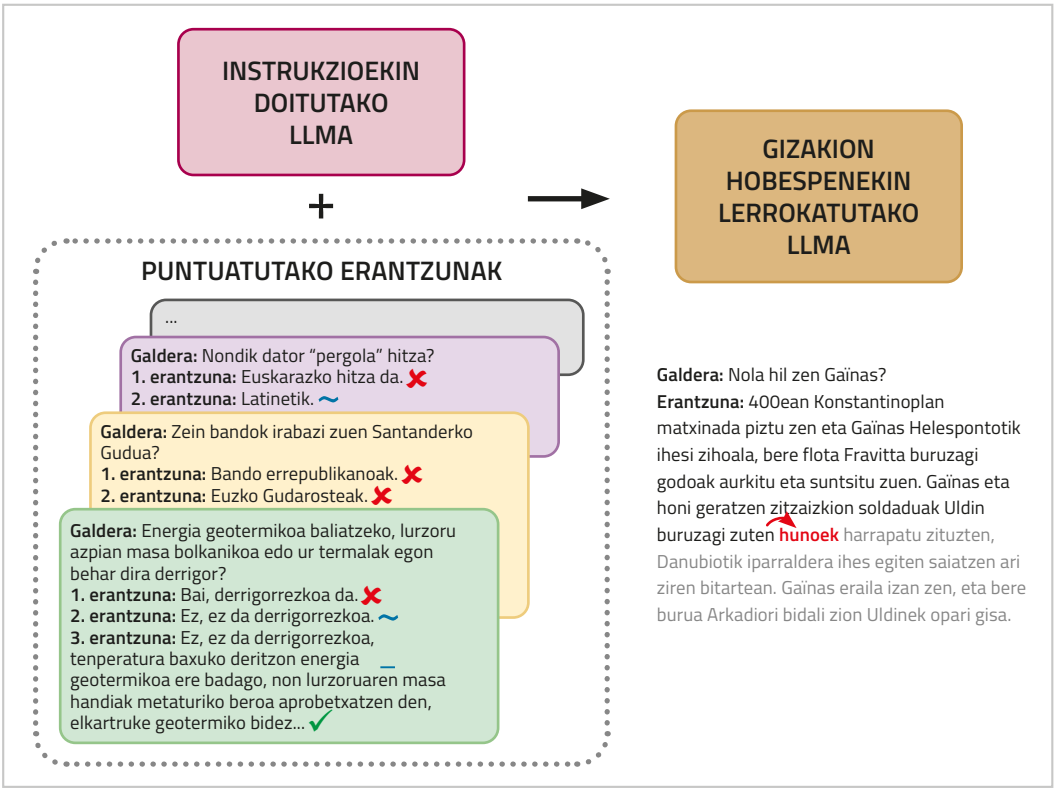
> *"Hizkuntza guztiak nahiko ongi ikastera irits daiteke LLM bat, transferentzia bidezko ikasketa deritzon propietatea dela eta"*

izan du, eta euskaraz oso ongi egiten duela onartu behar da.

Hori horrela izanik ere, hainbat arrazoirengatik (burujabetza teknologikoa, gure hizkuntzaren etorkizuna AEBko multinazionalen esku ez egotea, erraldoi teknologikoen tresnak erabiltzeak dakarren pribatutasunaren galera...), komenigarria da euskarazko LLM edo txatbotak berton garatzea, eta horretan ari gara Euskal Herriko zenbait erakunde.

Euskarazko LLM bat zerotik eraikitzea, hala ere, ez da lan erraza. Ongi funtziona dezaten, milioika milioi hitzekin entrenatu behar izaten dira, eta euskaraz ez dago hainbeste inondik ere. Itzulpen automatikoa baliatuta lor daiteke testu-kopuru nahikoa, eta horrekin ere proba pilotuak egin ditugu, baina emaitzak ez dira nahikoa onak izan. Gainera, sare erraldoi horien entrenamendu osoak makina oso ahaltsuak behar ditu, eta ikaragarri luze jotzen

3. FASEA. GIZAKION HOBESPENEKIN LERROKATZEA



du. Hori dela-eta, oso garestia da, eta, beraz, ez da bideragarria.

Horregatik, baliatu ohi den bidea da aurre-entrenatutako LLM fundazional libre bat hartu eta horri doikuntza fina edo *fine-tuning* egitea, euskaraz hobeto ikas dezan. Finean, doikuntza hori da entrenamenduarekin jarraitzea euskarazko testuak baliatuta, eta horregatik aurre-entrenamendu jarraitua edo *continual pre-training* ere deitzen zaio. Kontua da aurre-entrenatutako LLM horrek jada badakizkiela beste hizkuntza batzuk eta jakintza orokorra eta beste gaitasun batzuk badituela, eta, transferentzia bidezko ikasketari esker, euskara ikasteko ez dituela hainbeste testu edo entrenamendu-denbora behar. Gainera, nahi izanez gero LoRA (*Low-Rank Adaptation*) moduko teknikak baliatu daitezke sare osoaren ordez zati bat soilik doitu behar izateko, eta memoria-eskakizunak asko murrizten dira horrela.

Bide horretako azken aldiko lanetan, Meta enpresak garatutako eta lizentzia librean utzitako LLaMa erabiltzen da. EHUko HiTZ zentroak, adibidez, LLaMa 2 oinarri hartu eta EusCrawl testu-bildumarekin entrenamendua jarraituta, Latxa euskarazko LLMa atera zuen iazko urtarrilean, zeina gero, apirilean, testu-corpus handiago batekin egokitu eta hobetu baitzuen. Latxak beste edozein LLMk baino emaitza hobeak lortu zituen euskarazko gaitasun-probatan (EGAko atariko probekin ebaluatu zen), eta galdera edo ataza orokorretan GPT 4k soilik gainditzen zuen (galdera orokorretarako, irakurketa-ulermenezko galderetarako eta oposizioetako galderetarako prestatu zituzten ebaluaziorako multzo batzuekin).

Elhuyarreko gure Orai NLP Teknologiak zentroan, LLaMa 3.1 hartu genuen eta ZelaiHandi corpusarekin doitu (ZelaiHandi Oraik bildutako euskarazko testu libreen bilduma handiena da, 521 milioi hitzekoa), eta emaitza Llama-eus-8B da, zeina iazko irailean aurkeztu baitzen. Euskarazko eredu fundazional arinen artean (10 mila milioi parametro baino gutxiagokoak) emaitza hoberenak lortu ditu ataza-mota guztietan, eta zenbait atazatan eredu askoz handiagoek baino emaitza hobeak ere ematen ditu.

> *"Euskarazko eredu fundazional arinen artean, Llama-eus-8B-k emaitza hoberenak lortu ditu ataza-mota guztietan"*

Eta euskarazko txatbotak?

Emaitza horiek ikusita, batek pentsa dezake jada badugula euskarazko ChatGPT moduko bat berton sortua. Baina ikusi dugunez, bertoko ereduek euskarazko ataza-eskaera mota askorentzat lortzen dituzten emaitzak oraindik ez dira iristen GPT4ren mailara; eta, batez ere, emaitzak ingelesarekin lortzen direnak baino nabarmen baxuagoak dira.

Horren arrazoia goian azaldu dugu: LLM batetik txatbot funtzional bat lortzeko, instrukzioen doikuntza fina eta gizakion hobespenekin lerrokatzeko entrenamendu-faseak ere behar dira. Ikusi da txatbot komertzialen errendimendu onaren arrazoi nagusietako bat bi urrats horietan erabilitako

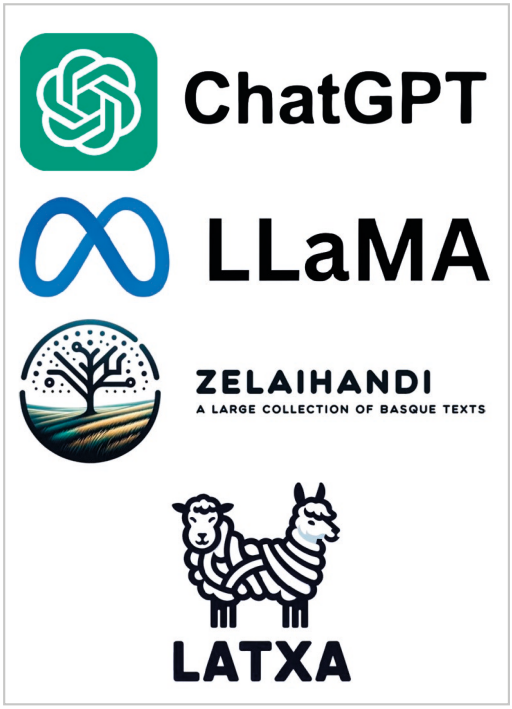
"Enpresa handien txatbotek alborapen kultural handia dute; euskal gaien galderen % 20 soilik asmatzen dute"

datu-sortak direla, oso handiak eta kalitatezkoak. Eta datu-mota horiek (eskaera-soluzio sortak eta pertsonek ebaluatutako erantzun-sortak) —aurre-entrenamendurako saretik nahiko automatikoki kopuru handian lor daitezkeen testuak ez bezala— pertsonek eskuz sortu beharrekoak dira, eta hori oso garestia da. Erraldoi teknologikoek diru eta eskulan asko esleitzen diote horri, eta datu-sorta horiek ez dituzte libre uzten. Euskaraz horrelakoak sortzeko ez dago baliabide nahikorik. Horrelako datu-sorta libre batzuk egon badaude, baina ez dira behar bezain handiak, ez daude euskaraz... Instrukzioen doikuntza finik gabe ere lor liteke ataza jakin batzuk hobeto egitea galderan bertan egin nahi den atazaren adibide bat edo batzuk emanik *(prompt engineering* edo *in-context learning* deritzo teknika horri), baina emaitzak ez dira doikuntzarekin lortzen direnak bezainbestekoak.

Hori guztia dela eta, oraindik lan handia dago egiteko gure txatbot funtzional propioa izateko. Horrek ez du esan nahi horretan ari ez garenik. Orai zentroan, esaterako, martxan ditugu zenbait lan. Adibidez, euskaraz jakiteaz harago eta euskaraz eskaera orokorrei erantzuteaz harago, LLMek eta txatbotek Euskal Herriari eta euskal kulturari buruzko ezagutza bat ote duten aztertu dugu, hori ebaluatzeko datu-sorta bat sortuz eta libre jarriz. Datu-sorta ingelesez dago (azken finean, enpresa handien txatbotak ingelesez funtzionatzen dute hobe, eta haien euskal gaiei buruzko ezagutza ebaluatu nahi zen, ez euskarazko gaitasuna), eta ingelesez egin zaizkie galderak txatbotei. Ondorioa izan da alborapen kultural handia dutela, eta euskal gaien galderen % 20 inguru soilik asmatzen dutela batez beste. Eredu fundazional libreetan hainbat teknika probatuz, hori hobetzen saiatu gara, eta arrakastatsuak izan dira saioak, asmatze-tasa % 80 ingururaino igo baita. HiTZ zentroak ere egin du antzeko lan bat eta ebaluazio bat.

Euskarazko LLMen alborapena ere badugu langai Orai zentroan. LLMen alborapenak neurtzeko erabiltzen den BBQ datu-sorta euskaratu eta euskal testuingurura egokitu dugu, eta publiko egin BasqBBQ izenarekin eta lizentzia librearekin. Hori baliatuta, euskarazko LLMek (Latxa eta LLaMa-eus-8b) dituzten alborapenak neurtu dira, eta horiek oinarrian duten jatorrizko LLaMa ereduarenekin alderatu. Eta ikusi da euskarara egokitutako ereduek ez dutela alborapen handiagoa, alderantziz baizik.

Azkenik, instrukzioen doikuntza finaren eta gizakion hobespenekin lerrokatzearen inguruko lehenengo esperimentuak ere egin ditugu. Horretarako, libreki eskuragarri dauden ingelesezko zenbait datu-sorta eskuratu ditugu, bai instrukzioenak (eskaera-soluzio motakoak) bai puntuatutako erantzunenak, eta, itzulpen automatiko bidez euskarara itzuli ondoren, beste bi entrenamendu-fase horiek pasatu dizkiogu LLaMa-eus-8B gure eredu fundazionalari. Hori eginda, entrenamendu-fase guztietatik pasatutako lehenengo euskarazko txatbota eraiki dugu. Eta emaitzak konparatu ditugu Metak, bere datu pribatuak baliatuta, instrukzioekin doituta eta gizakion hobespenekin lerrokatuta ateratako antzeko tamainako LLaMa txat-ereduarenekin, eta ikusi dugu gureak askoz hobeto funtzionatzen duela euskarazko sorkuntza-ataza horietan. Hala ere, emaitzen kalitatea oraindik ez da iristen eredu horiek ingelesezko atazetan lortzen dutenera, eta gutxiago ChatGPT bezalako eredu itxienera. Azken finean, esan bezala, instrukzio- eta DPO-



entrenamendurako datu-sorta irekiak ez dira erraldoi teknologikoek dituztenak bezain handiak, eta, gainera, itzulpen automatikoz itzuli izanak ere badu nolabaiteko eragina.

Lan horiek guztiak eta horrelako beste asko egin beharko dira oraindik euskarazko txatbot funtzional orotariko bat izateko. Baina horrelako bat lortzen denean ere, euskal gizarteak erabiltzeko moduan eskuragarri jartzea arazoa suerta daiteke. Izan ere, sare neuronal erraldoi horiek entrenatzea garestia den bezala, erabili ahal izateko martxan izatea ere oso garestia da, makina oso ahaltsuak behar baitituzte. Txatbot komertzial arrakastatsuenak eskaintzen dituzten erraldoi teknologiko estatubatuarrak iraultza honen aurrealdean egoteagatik eta merkatu-kuota irabazteagatik ari dira, baina dirukopuru izugarria galtzen ari dira. Eredu horiek ez dira errentagarriak, eta, bestalde, hor dago horrelako makina handiek ingurumenari egiten dioten kaltea ere. Horregatik, txatbot hauen guztien emaitzak hobetzea bezain garrantzitsua da optimizazioa, hau da, emaitza berak lortzea sare neuronal ekonomikoki eta ekologikoki jasangarriagoak baliatuta.

Eta justu bide horretatik omen doa urtarrilaren amaieran ezaguna egin zen Txinako DeepSeek enpresaren kode irekiko LLMa, DeepSeek-V3. Berau erabiltzeko app-aren deskargek ChatGPT-renak gainditu zituzten AEBtan, harenekin konparagarriak diren emaitzak eskaintzen baitzituen askoz prezio merkeagoan. Horrek lurrikara eragin zuten AEBtako AA eta txip enpresen burtsako kotizazioetan eta etorkizuneko asmo eta esperantzetan. Baina emaitzen kalitate handiaz gain, kontua da Txinari AEBek jarritako txipen enbargoak behartuta, LLMak txip ez hain ahaltsuetan garatu ahal izateko bideak bilatu behar izan dituztela eta, antza, asmatu dutela.

LLMaren egituran eta entrenamenduan besteek orain arte erabili ez dituzten zenbait aldaera baliatuta, DeepSeek-V3-ren entrenamenduaren kostua GPT-4-renaren % 6 izan dela dio enpresak, eta LLaMa 3.1-en entrenamenduaren energiaren % 10 soilik behar izan duela. Eredua zerbitzu gisa eskaintzearen kostua ere askoz txikiagoa da eta, hala, merkeago eskaini arren, dirua galtzen ez duen bakarra ei da. Hau da, DeepSeek-en emaitzak ChatGPT bezalako eredu itxienen parekoak dira, LLaMa bezalako kode irekikoenak baino hobeak, kode irekikoa da, eta beste ereduen energia-beharraren (eta, beraz, kostuaren) hamarrena baino gutxiago omen du. Laster ikusiko da ea bide horren egokitasuna konfirmatzen den, beste ereduen garatzaileek ere bide hori hartzen duten, eta guretzat ere baliagarri den euskarazko eredu propioak azkarrago eta merkeago garatu eta eskaini ahal izateko. ●