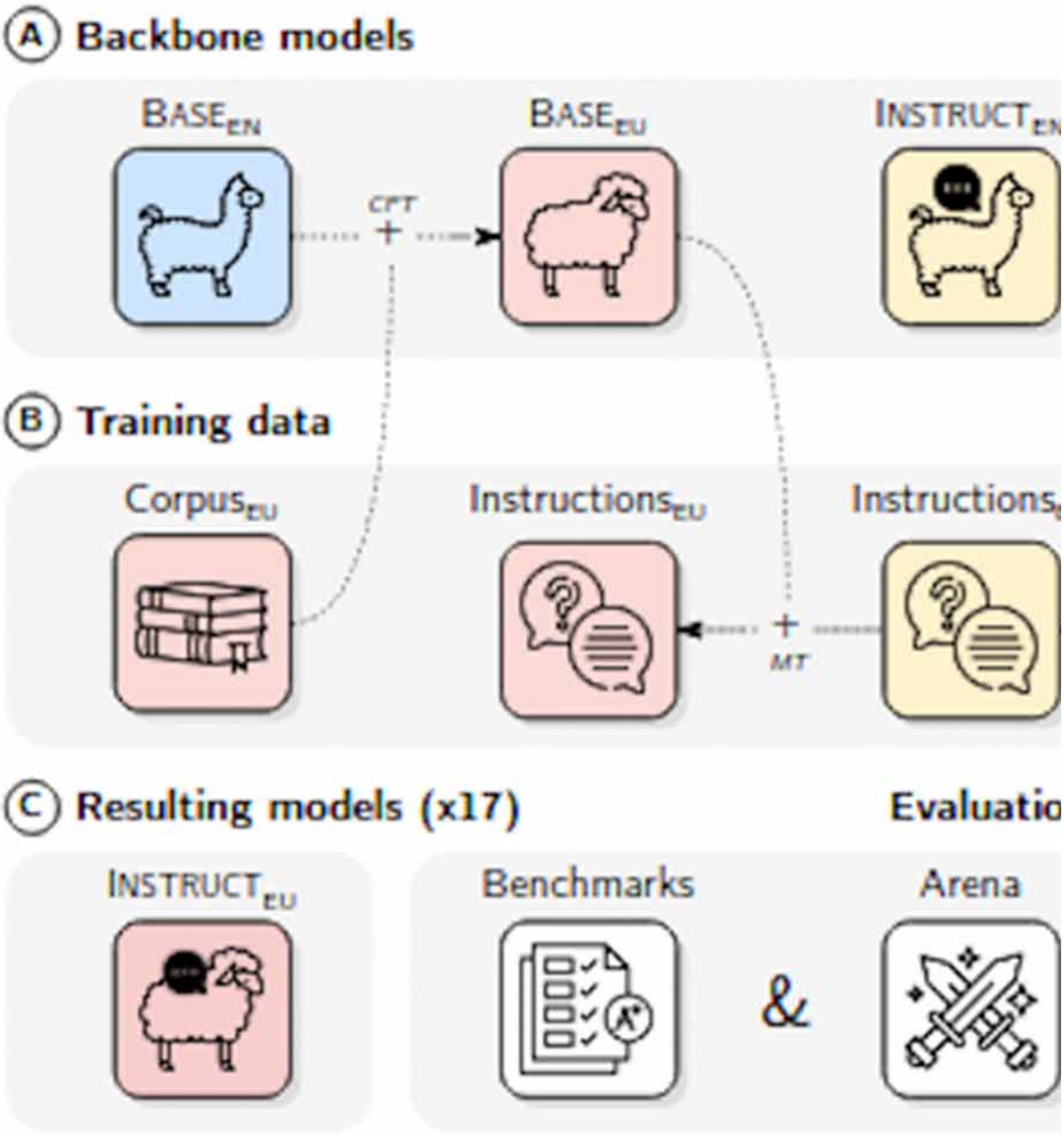
HiTZ zentroak hizkuntza txikiagoentzat txatb
modu berria deskubritu du

Elkarrizketak edukitzeko gai den euskarazko txatbotak egiteko modu berriaren berri eman

ALBISTEAK 2025/06/18



ChatGPT bezalako txatboten erabilera etengabe handitzen ari da, eta lanbide askotan tresna ezir Haien gaitasunak hedatu eta hobetzen dihardute, hala nola, testuak laburtu, edozein gairi buruzk jasak egin, programatu, dokumentuak sortu edo dokumentu osoak hizkuntza batetik bestera itzul errendimendu parekoa duten bitartean, euskara bezalako hizkuntza txikiagoetara joatean kalitate hizkuntza horietarako dauden testu masak hain ugariak ez direlako. Adibidez, Internet irekian dag dagokionean, ingelesez euskaraz baina 1000 aldiz dokumentu gehiago daude, eta gaztelaniaz e gehiago. Zientzialarion artean galdera irekia izaten jarraitzen du ea posible den halako testu mas pareko emaitza onak lortzea. Eneko Agirre EHUko HiTZ ikerketa zentroaren zuzendariaren uste,

aurrerapen handia egin da galdera horren baiezko erantzunaren bidean, baina oraindik inkognita

Euskarazko txatbota eraikitzeko pausoak

HiTZ zentroak duela urtebete inguru publikatu zuen euskarazko hizkuntza-eredu irekia egiteko m hoberenaren saria irabazi zuen arlo honetako kongresu entzutetsuenean. Latxa izeneko hizkuntz kodearekin batera, libre utzi zituen, enpresa eta erakunde interesatuak erabili zitzaten beraien ga egindako ebaluazio estandarizatuetan orduko ChatGPTren pareko ezagutza zuela erakutsi zuten euskararen ezagutzari buruzko frogetan hobeto ari zela ere. Latxa garatzen duen lantalde zabala dio “euskara bezalako hizkuntza batentzat lehen aldia izan zen halako kalitatea lortzen, azken fin erraldoi teknologikoekin lehiatzen aritu ginen eredu irekietan oinarrituta gainera”.

Naiara Pérez ere garatzaileetako bat da, eta kontatzen duenez duela urtebete eraikitako Latxak “ baina ez zen gai elkarrizketak edukitzeko, ez eta ataza konkretuak egiteko aginduak jarraitzeko. H nahiz eta erantzuteko adina ezagutza jakin, ez zen konturatzen erantzuna sortu behar zuela. Gau aginduak edo beste edozein agindu emanez gero. Modu berean, testu toxikoak edo desegokiak s inongo filtrorik halakoak ez egiteko”.

Metodo berria

Artikulu zientifiko berri honetan HiTZeko kideek elkarrizketak edukitzeko gai den euskarazko txatl berriaren berri eman dute. Artikulu zientifikoa errebisiopean dago, eta onarpenaren zain dagoen l dute. Abiapuntua Meta enpresako ikerkuntza zentroak eraikitako hizkuntza-eredu ireki eleaniztun “Eredu libreen artean ahaltsuenetakoa izanda ere euskaraz trakets ari da, eta helburua euskaraz eraikitzea da. Bide ohikoena Llama hartu, euskarazko testu-masekin elikatu, eta ondoren euskara eta erabilera toxikoak bereizteko adibideekin elikatzea da. Tamalez, kalitate handiko txatbota erai erabilera adibide asko sortu behar dira, eta horrek oso garestia den eskulana eskatzen du. Enpre ahal izan dute oraingoz”, dio Eneko Agirrek.

HiTZeko kideek hainbat bide probatu dituzte eskuzko lan hori ekiditeko, eta uste zenaren kontra, efizientea topatu dute eskuzko lanik gabe kalitate handiko txatbota euskarara egokitzeko. Egindal garapenek erakusten dute nahikoa dela Llama txatbota euskarazko testu-masarekin entrenatzen horretarako gakoa da ahanztura katastrofikoa (catastrophic forgetting ingelesez) deritzan arazoar aplikatzen asmatzea.

Zein da aukera hoberena? Ebaluatoia

Baina nola ebaluatu “edozer gauza” egin dezaketen hizkuntza-eredu hauek? Ebaluazioa egiteko gizarteari laguntza eskatu zioten duela hilabete batzuk. *Ebaluatoia* deitu zen ekimenean, parte-ha edo eskaera bat bota, bi txatboten erantzuna jaso, eta erantzun gustukoenaren alde bozkatzeko a hamahiru mila hobespen lortu ziren. “Probatu ziren txatboten artean Latxaren aldaera ezberdinak esker frogatu da HiTZek asmatutako metodoa euskarazko txatbotak garatzeko metodo hoberena erraldoiko Latxa 8B tamainakoa baina dezente hobea dela. Latxaren aldaera handiena, hain zuze diren OpenAIren GPT-4o eta Anthropic-en Claude-etik oso gertu dagoela ikusi da”.

Latxa txatbota garatzaileentza eskuragarri

Egindako lanak bide berriak irekitzen ditu. Alde batetik metodoa bera Llama baino txatbot ireki ah

daiteke, etorkizunean Latxa ahaltsuagoak eraikitzea ahalbideratzen. Gainera, metodoa euskara e aplikatu daiteke, kopuru antzekoa duten testu-masak erabiliaz.

Bestetik, Latxa eredu guztiak eskuragarri daude, garatzaile eta enpresek beraien beharretara ego doan. Eusko Jaurlaritzaren Hizkuntza-Politikarako Sailburuordetzarekin elkarlanean, Euskal Herr nahi dutenerako erabili dezaten. Ehun kidetik gora erabili dute bi astez, eta Latxaren indargune e Tartean galdera zehatza egin zaie: “Latxa jendarteak erabili dezan prest dago?” Erantzuna baiezk batzuetan hobekuntza batzuk iradoki diren. Egun HiTZ zentroa publiko zabalaren esku uzteko mo aztertzen ari da Eusko Jaurlaritzarekin batera.

“Lan hau mugarria izanda ere, Latxa hobetu eta hedatzen lan asko dago egiteko. Txatbotak eteng gaitasun berriekin, eta edonon erabiltzen ari dira. Google bera bilatzailea txatbot bihurtzeko bidea dago gailu hauek munduaren ikuspegi konkretuak islatzen dituztelako, eta horrek hegemonia ling dezakeelako. Euskararen kasuan, hizkuntzaren erabileran eduki dezakeen inpaktuaz haratago, b ikuspegiak modu egokian islatzea ere jokoan dago”, azpimarratu du HiTZ zentroko zuzendariak.

Informazio osagarria

Artikulua, Latxa eredu guztiak, erabilitako corpusak eta proba bankuak hemen eskura daitezke: h zentroa/latxa-instruct. Latxa ereduak “Llama License” izenekoa hartzen dute oinordetzan. Eredu l merkataritza jarduera ahalbidetzen ditu.

HiTZetik eskerrik beroena eman nahi diote Ebaluatoian lagundu duten pertsona guztiei, ikerketa l ezinbestekoak izan dira eta. Azpimarratzekoa da ere Latxa entrenatu ahal izateko euskarazko tes partekatzen dituzten eragileen ekarpena (Tokikom, Berria, Hitza, Wikipedia, Argia, Bilbao Hiria Irr beste). “Esku zabaltasun hori ezinbestekoa da euskarazko txatbotak eraikitzeko, eta HiTZen parte nahi zaie, baita Egunkariari ere”.

Latxa Eusko Jaurlaritzak finantzatutako IKER-GAITU proiektuaren esparruan garatu da. Eraldake Publikorako Ministerioak eta Suspertze, Eraldatze eta Erresilientzia Planak, Europar Batasuneko finantzatu du, ILENIA proiektuarekin lankidetzan, 2022/TL22/00215335 erreferentzia duena. HiTz konputazio azpiegitura (HPC) propioa erabili du, eta azken ereduak CINECAren Leonardo superc dira, EuroHPC Joint Undertaking barruan (EHPC-EXT-2024E01-042 proiektua).

irudi galeria

partekatzea