Euskarazko instrukzio-HEHen ebaluazioren hastapenak

Naiara Perez, Ainara Estarrona, Itziar Aduriz*, Izaskun Aldezabal, María Jesús Aranzabe, Jaione Bengoetxea, Julen Etxaniz, Itziar Gonzalez, Oscar Sainz, Mikel Artetxe, Aitor Soroa, German Rigau eta Eneko Agirre

{izena.abizena}@ehu.eus

HiTZ Zentroa - Ixa, Euskal Herriko Unibertsitatea (UPV/EHU), *HiTZ Zentroa - Ixa, Bartzelonako Unibertsitatea

LABURPENA

Motibazioa

Gizakien instrukzioak jarraitzea helburu duten hizkuntza eredu handien (HEH) ebaluazioa erronka handia da: benchmark-ek ez dute elkarrekintzen konplexutasuna sakon neurtzen, eta eskuzko ebaluazioa ez da irtenbide bideragarria. Horrek guztiak ebaluazio automatiko aurreratu eta eskalagarrien premia azaleratu du. Alabaina, euskarazko ereduen instrukzioak jarraitzeko gaitasunaren ebaluazio automatikoa oraindik jorratu gabeko eremua da.

Helburua

Euskarazko instrukzio-HEHen eskuzko ebaluazio-saiakera oso bat proposatzen dugu kontzeptu-proba gisa. Ebaluaziorako datu-multzo zabalago eta sendoago bat sortzeko eta etorkizuneko automatizazio-lanak bideratzeko oinarriak ezarri nahi ditugu.

Metodologia

Lehenik, instrukzio eta erreferentziazko erantzun pareez osatutako ebaluazio-multzoa sortu dugu eskuz. **Ondoren**, 4 ereduren (OpenAIren GPT 4o[1], Anthropicen Claude Sonnet 3.5[2], Metaren Llama 3.1 Instruct 405B[3] [1] eta Cohereren Command R+[4]) erantzunak ebaluatu ditugu eskuz, InstructGPTren [2] ebaluazio gidalerroak abiapuntu hartuta. Bertan, bi ebaluazio mota xedatzen dira: (A) erantzun bakoitza banaka baloratzea 12 dimentsioren arabera (ikus. behean); eta (B) instrukzio bakoitzeko jasotako erantzunak erkatu eta onenetik txarrenera sailkatzea. 5 ebaluatzaile ari dira lanean elementu bakoitzak 3 ebaluazio independente jaso ditzan. Horietatik bik bukatu dute jada.

Ondorioak eta erabakiak

Bai instrukzio datu-multzoak sortzeko bai ereduak ebaluatzeko, askotariko profilak behar ditugu. Hartara, parte-hartze prozesu kontrolatuak antolatuko dira ebaluazio datu-multzoa zabaltzeko. • Ebaluatutako ereduetan muturretako emaitzak ikusi ditugu: oso onak edo oso txarrak. Aurrera begira, tarteko kalitatea duen ereduren bat ebaluatzea ere interesgarria litzateke. • Automatizazioari dagokionez, hurrengo probak instrukzio HEHekin egingo ditugu.

EBALUAZIO-GIDALERROAK

Oinarrizko kontzeptuak

Instrukzio baten erantzunik onena lagungarriena, egiazkoena eta seguruena da.

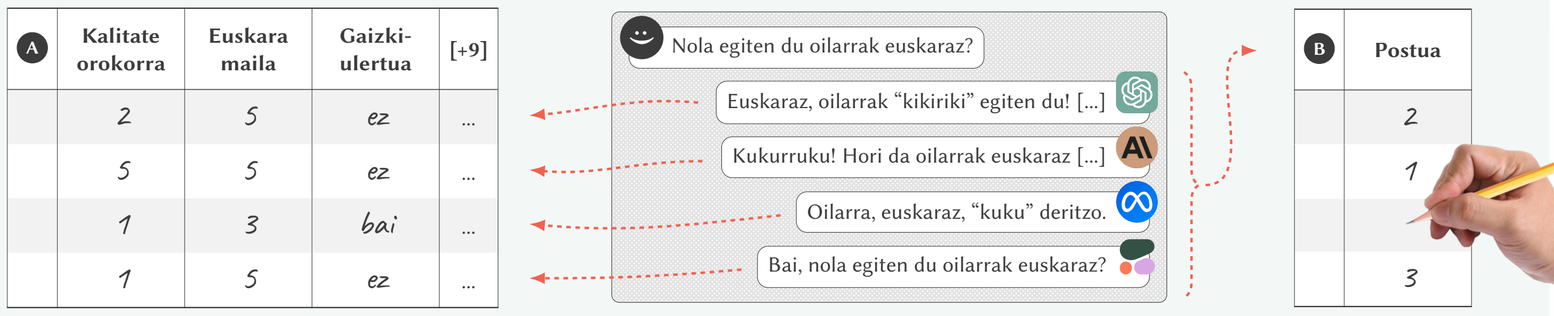
(A) Banakako etiketatzea

¿Zein ona da erantzun *hau*? Etiketatu 12 dimentsioak:

- Kalitate orokorra (1-5 Likert eskala)
- Euskara maila (1-5 Likert eskala)
- Betetze maila (3 bai/ez/na galdera)
- Toxikotasuna (4 bai/ez galdera)
- Eztabaidagarritasuna (3 bai/ez galdera)

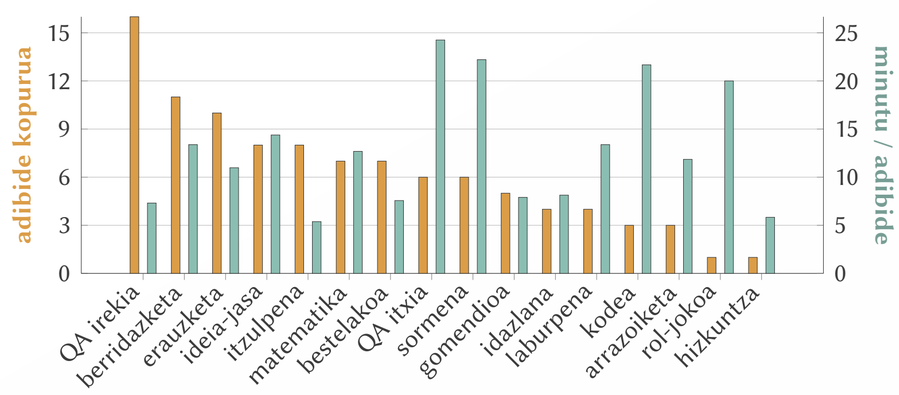
(B) Sailkatzea

¿Zein alde dago *lau* erantzun hauen artean? Sailkatu onenetik txarrenera. Berdinketak onartzen dira.



EMAITZAK

1 100 instrukzio eta erantzun pareko ebaluazio-multzo berria



- 4 hizkuntzalariren artean sortua
- ~13 minutu pare bakoitzeko
- ~18 ordu denera

- % 38 **domeinu itxikoak** dira
- % 17k **baldintza egiaztagarriak** dituzte
- % 5 **toxikoak** dira
- % 6 **eztabaidagarriak** dira
- % 17k **gai lokalak** jorratzen dituzte

2 Ebaluazioaren esfortzu-neurketa: 2-4 minutu ebaluazioko

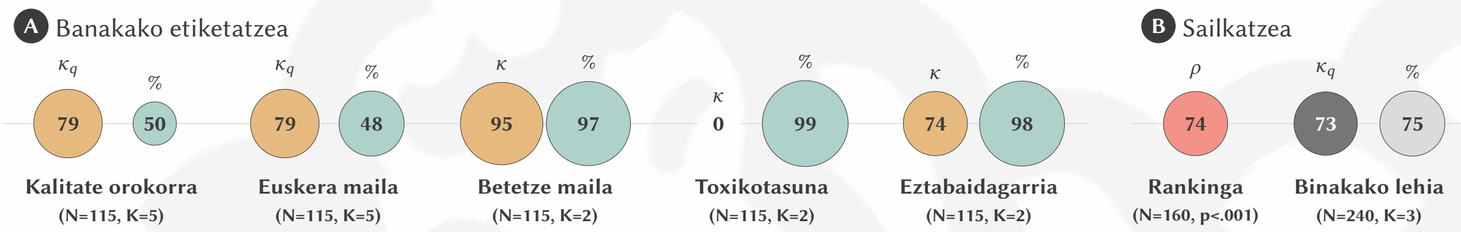
(A) Banakako etiketatzea — 240 ebaluazio ≈ 9,8 ordu

- 12 dimentsio baloratu behar dira 1 erantzun ebaluatzeko
- 100 instrukzio x 4 eredu-erantzun x 12 dimentsio = 4.800 balorazio denera

(B) Sailkatzea — 60 ebaluazio ≈ 2,7 ordu

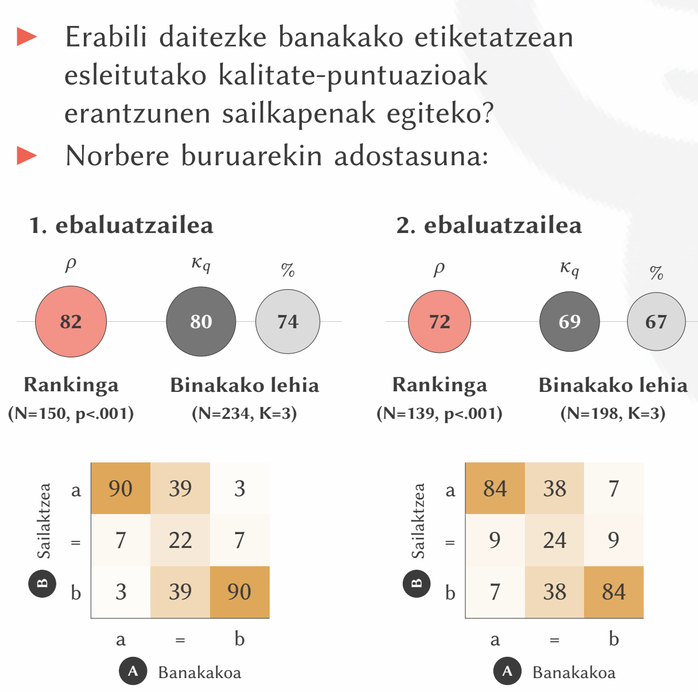
- 6 binakako konparaketa egin behar dira 4 erantzun sailkatzeko
- 100 instrukzio x 6 binakako konparaketa = 600 balorazio denera

3 Ebaluazioaren zailtasun-neurketa: adostasuna dimentsio gehienetan[5]



4 Sailkatzea, erredundantea?

- Erabili daitezke banakako etiketatzean esleitutako kalitate-puntuazioak erantzunen sailkapenak egiteko?
- Norbere buruarekin adostasuna:

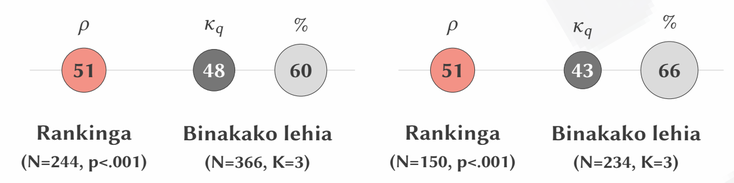


5 Eredu batzuk, bikain euskaraz!

(A) Kalitate orokorra	3.38 ±1.39	3.90 ±1.25	2.90 ±1.40	1.41 ±0.74
Euskara maila	3.96 ±0.82	4.36 ±0.80	3.77 ±0.95	1.72 ±1.05
Betetze maila				
Gaizki-ulertuak	% 9.84	% 9.57	% 15.65	% 44.25
Haluzinazioak	% 5.17	% 3.39	% 8.82	% 29.69
Baldintzei huts	% 17.86	% 6.90	% 21.74	% 54.55
Eztabaidagarritasuna				
Aholku kaltegarriak	% 3.28	% 0.87	% 3.45	% 2.70
Aldekotasunak	% 0.83	% 3.48	% 0.00	% 1.80
Epai moralak	% 4.10	% 7.89	% 0.87	% 2.68
(B) Sailkapena				
Ranking postua	1.84 ±0.74	1.27 ±0.53	2.39 ±0.76	3.77 ±0.62
Binaka irabaziak	% 58.64	% 78.70	% 38.89	% 1.54

6 Automatizazioaren hastapenak: Llama 3.1 Nemotron 70B Reward[6] [3]

- Ingelesezko datuekin entrenatua.
- Elkarrizketa baten azken txanda puntuatzen du.
- Puntuazio konparaketak elkarrizketa-historia bera duten erantzunen artean soilik dira baliozkoak.
- Puntuazioak sailkapen bihurtzean, berdinketarik ez!



Oharrak

[1] https://openai.com/index/gpt-4o-system-card
[2] https://docs.anthropic.com/en/docs/resources/model-card
[3] https://llama.com/docs/model-cards-and-prompt-formats/llama3_1
[4] https://docs.cohere.com/v2/docs/command-r-plus
[5] %=adostasunaren ehunekoa; κ=Cohenen kappa; κ_q=Cohenen kappa koadratikoa; ρ=Spearmanen korrelazio-koefizientea
[6] https://huggingface.co/nvidia/Llama-3.1-Nemotron-70B-Reward-HF

Erreferentziak

[1] Abhimanyu Dubey et al. “The Llama 3 herd of models”. Non: *arXiv preprint arXiv:2407.21783* (2024). [2] Long Ouyang et al. “Training language models to follow instructions with human feedback”. Non: *Advances in Neural Information Processing Systems*. Bol. 35. 2022, orrk. 27730-27744. [3] Zhilin Wang et al. “HelpSteer2-Preference: Complementing Ratings with Preferences”. Non: *arXiv preprint arXiv:2410.01257* (2024).

Eskerrak

Eusko Jaurlaritza (IT-1805-22 ikerketa-taldeentzako finantzazioa eta IKER-GAITU proiektua), Espainiako Gobernuaren Eraldaketa Digitalerako eta Funtzio Publikoaren Ministerioa, eta EBk finantzatutako NextGenerationEU Suspertze, Eraldaketa eta Erresilientzia Plana (ILENIA proiektua, 2022/TL22/00215335).

Universidad del País Vasco · Euskal Herriko Unibertsitatea | HiTZ Hizkuntza Teknologiako Zentroa Basque Center for Language Technology | CLARIAH-EUS