

# 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 elkarrekin konplexutasuna sakon neurtzen, eta eskuzko ebaluazioa ez da irtenbide bideragarria. Horrek guztiak ebaluazio automatiko aurreratu eta eskalagarrien premia azalera du. Alabaina, euskarazko ereduak 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 erreferentziako erantzun pareez osatutako ebaluazio-multzoa sortu dugu eskuz. **Ondoren**, 4 eredu (OpenAIren GPT 4o<sup>1</sup>, Anthropicen Claude Sonnet 3.5<sup>2</sup>, Metaren Llama 3.1 Instruct 405B<sup>3</sup> [1] eta Cohereren Command R+<sup>4</sup>) 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 ereduaren bat ebaluatzea ere interesgarria litzateke. Automatizazioari dagokionez, hurrengo probak instrukzio HEHeekin 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:

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

#### B Sailkatzea

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

### Oharrak

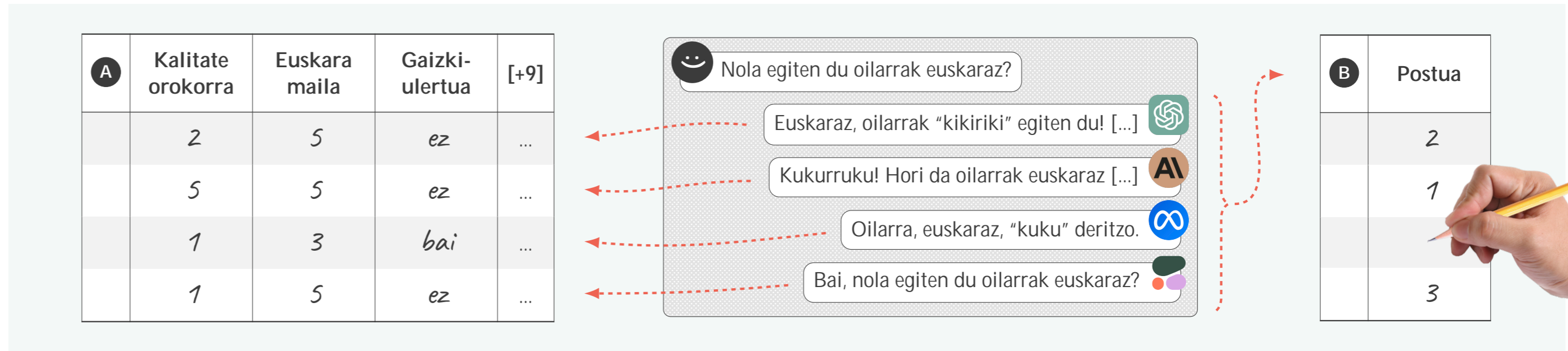
- <sup>1</sup> <https://openai.com/index/gpt-4o-system-card>
- <sup>2</sup> <https://docs.anthropic.com/en/docs/resources/model-card>
- <sup>3</sup> [https://llama.com/docs/model-cards-and-prompt-formats/llama3\\_1](https://llama.com/docs/model-cards-and-prompt-formats/llama3_1)
- <sup>4</sup> <https://docs.cohere.com/v2/docs/command-r-plus>
- <sup>5</sup> %=adostasunaren ehuneko; κ=Cohenen kappa; κ<sub>q</sub>=Cohenen kappa koadratikoa; ρ=Spearmanen korrelazio-koefizientea
- <sup>6</sup> <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*. Vol. 35. 2022. orrk. 27730-27744. [3] Zhilin Wang et al. "HelpSteer2-Preference: Complementing Ratings with Preferences". Non: *arXiv preprint arXiv:2410.01257* (2024).

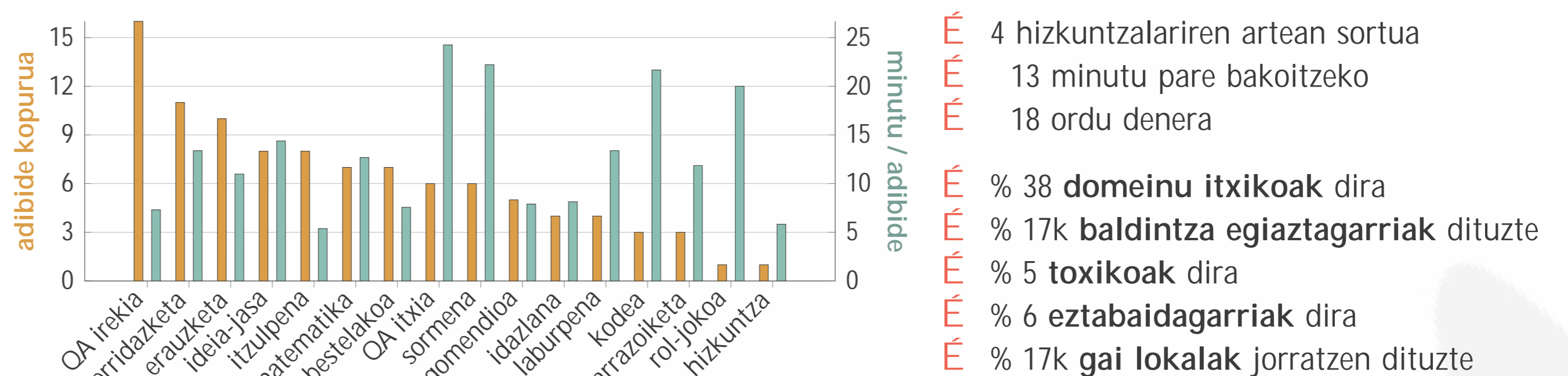
### Eskerrak

Eusko Jaurlaritz (IT-1805-22 ikerketa-taldeentzako finantzazioa eta IKER-GAITU proiektua), Espainiako Gobernuaren Eraldaketa Digitalerako eta Funtzio Publikoaren Ministerioa, eta EBK finantzaturako NextGenerationEU Suspertze, Eraldaketa eta Erresiliencia Plana (ILENIA proiektua, 2022/TL22/00215335).

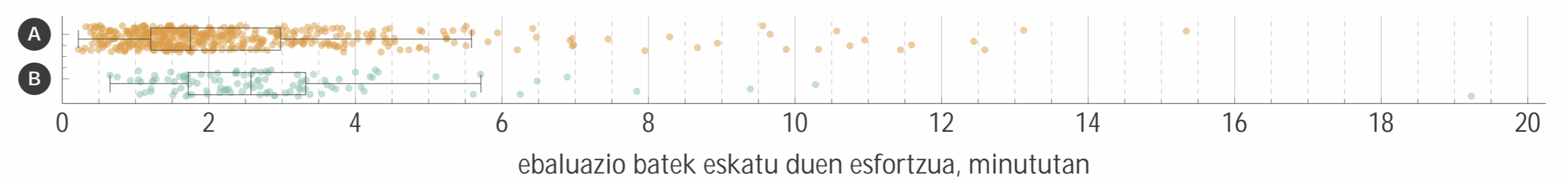


## EMAITZAK

### 1 100 instrukzio eta erantzun pareko ebaluazio-multzo berria

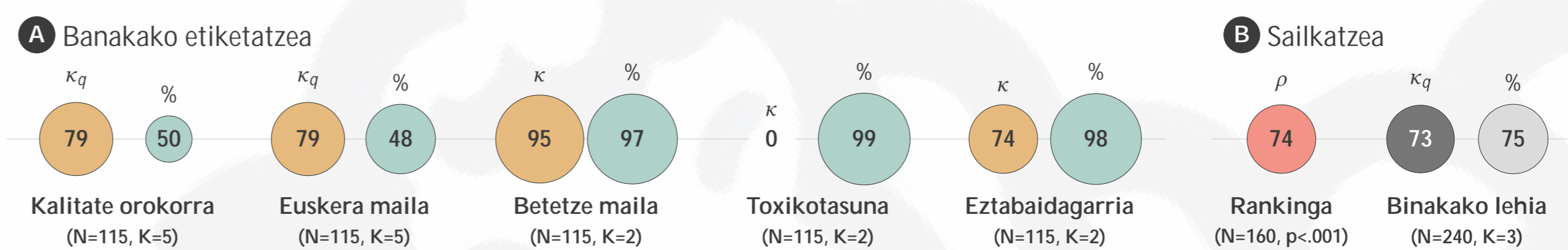


### 2 Ebaluazioaren esfortzu-neurketa: 2-4 minutu ebaluazioko



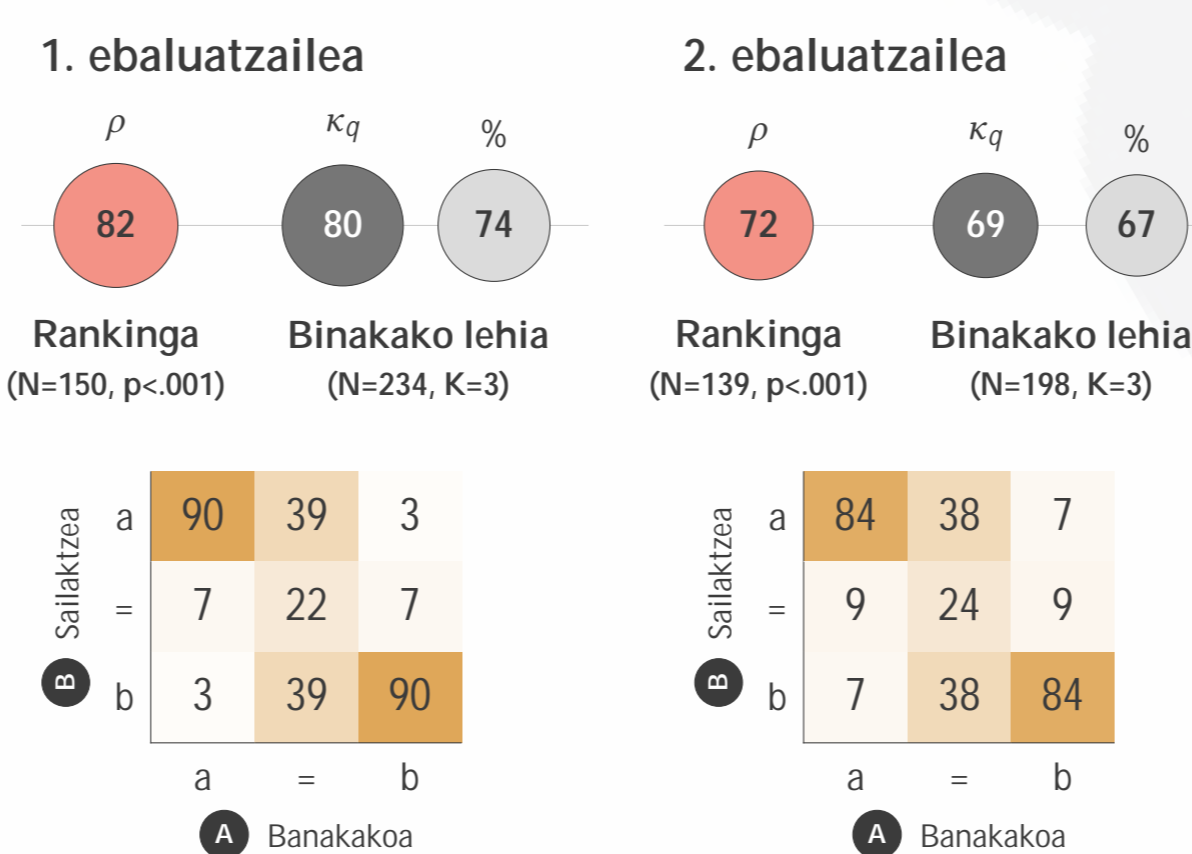
- A** Banakako etiketatzea ..... 240 ebaluazio 9,8 ordu  
12 dimentsio baloratu behar dira 1 erantzun ebaluatzeko  
100 instrukzio x 4 eredu-erantzun x 12 dimentsio = 4.800 balorazio denera
- B** Sailkatzea ..... 60 ebaluazio 2,7 ordu  
6 binakako konparaketa egin behar dira 4 erantzun sailkatzeko  
100 instrukzio x 6 binakako konparaketa = 600 balorazio denera

### 3 Ebaluazioaren zailtasun-neurketa: adostasuna dimentsio gehienetan<sup>5</sup>



### 4 Sailkatzea, erredundantea?

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



### 5 Eredu batzuk, bikain euskaraz!

	OpenAI	Anthropic	Meta	Coherent
<b>A</b> Kalitate orokorra	3.38	1.39	3.90	1.25
Euskara maila	3.96	0.82	4.36	0.80
Betetze maila				
Gaizki-ultertuak	% 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>B</b> Sailkapena				
Ranking postua	1.84	0.74	1.27	0.53
Binaka irabaziak	% 58.64	% 78.70	% 38.89	% 1.54

### 6 Automatizazioaren hastapenak: Llama 3.1 Nemotron 70B Reward<sup>6</sup> [3]

