Instructing Large Language Models for Low-Resource Languages: A Systematic Study for Basque

Oscar Sainz Naiara Perez Julen Etxaniz Joseba Fernandez de Landa
Itziar Aldabe Iker García-Ferrero Aimar Zabala Ekhi Azurmendi
German Rigau Eneko Agirre Mikel Artetxe Aitor Soroa
HiTZ Center - Ixa, University of the Basque Country UPV/EHU
{oscar.sainz,a.soroa}@ehu.eus

Abstract

Instructing language models with user intent requires large instruction datasets, which are only available for a limited set of languages. In this paper, we explore alternatives to conventional instruction adaptation pipelines in low-resource scenarios. We assume a realistic scenario for low-resource languages, where only the following are available: corpora in the target language, existing open-weight multilingual base and instructed backbone LLMs, and synthetically generated instructions sampled from the instructed backbone. We present a comprehensive set of experiments for Basque that systematically study different combinations of these components evaluated on benchmarks and human preferences from $1,680$ participants. Our conclusions show that target language corpora are essential, with synthetic instructions yielding robust models, and, most importantly, that using as backbone an instruction-tuned model outperforms using a base non-instructed model. Scaling up to Llama 3.1 Instruct 70B as backbone, our model comes near frontier models of much larger sizes for Basque, without using any Basque instructions. We release code, models, instruction datasets, and human preferences to support full reproducibility in future research on low-resource language adaptation.[1]

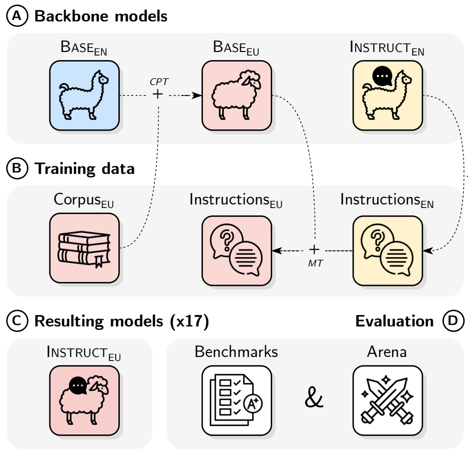


Figure 1: Systematic exploration of instruction-tuning strategies for low-resource languages. Our framework consists of: Ⓐ three backbone models (an existing base model, a continued-pretrained base model on the target language, and an existing instruct model); and Ⓑ different training data combinations, including target language corpora and synthetic instructions sampled and/or translated with the backbone models. We train Ⓒ experimental models from all possible combinations of these components, and perform Ⓓ comprehensive evaluation through both static benchmarks and human preferences to identify optimal adaptation paths.

1 Introduction

Large Language Models (LLMs), particularly open models, remain predominantly English-centric, with limited coverage for the vast majority of the world's languages. Despite recent efforts to incorporate additional languages during the pretraining of open LLMs, significant performance disparities still persist. Even the latest instruction-tuned models demonstrate markedly degraded capabilities when handling low-resource languages (Grandury et al., 2025). Critically, the English-focused nature of post-training processes has widened the performance gap between languages when comparing base and instruction-tuned models.

To overcome these limitations, open models can be adapted to new languages through continued training with limited resources (Etxaniz et al., 2024b). In the case of instruction-tuned models in particular, various efforts have emerged that typically follow a sequential approach (Ouyang et al., 2022): first adapting the base model through continued pretraining, then performing instruction tuning. While this multi-step process has become standard practice, little exploration has investigated alternative adaptation strategies. We question whether instruction-following capabilities could be directly

[1] github.com/hitz-zentroa/latxa-instruct

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 29125–29149
November 4-9, 2025 ©2025 Association for Computational Linguistics

transferred to new languages without dedicated instruction data, and whether instruction-tuned models could be adapted through continued pretraining, similar to base models.

Specifically, this work systematically explores diverse strategies beyond the conventional pipeline for developing instruction-tuned models for low-resource languages, seeking to identify optimal adaptation paths for Basque as our primary case study (see Fig. 1). We deliberately constrain our exploration to resources either readily available or creatable using open models, avoiding reliance on distillation from commercial state-of-the-art systems. While our investigation focuses on a single language, our findings likely generalize to many similarly-resourced languages: Basque represents an ideal test case, ranking approximately 50$^{\text{th}}$ in Common Crawl with a presence approximately $1,000$ times smaller than English,[2] and notably lacking pre-existing instruction datasets. This resource profile mirrors challenges faced by numerous other low-resource languages worldwide.

In addressing this research question, we further confront a key challenge in instruction-following LLM assessment: automated metrics often miss capabilities that matter to users (Chiang et al., 2024). Thus, we developed an evaluation framework combining traditional benchmarks with a crowdsourced LLM arena, where we mobilized the Basque-speaking community in a large-scale evaluation effort that gathered over $12,000$ preference annotations from $1,680$ participants. This initiative constitutes the largest human evaluation effort for a low-resource language to date.

Through this evaluation, our systematic exploration produced three key insights for developing instruction-tuned models in low-resource languages: **(1)** target language corpora is essential for performance—models lacking exposure to plain Basque text showed degradation regardless of other techniques; **(2)** while both monolingual and bilingual instruction datasets showed benefits, the latter produced consistent results across benchmark and human evaluations; and **(3)** starting from an instruction-tuned English model outperformed the approach of a base model learning to follow instructions, challenging the standard pipeline applied to low-resource languages.

In addition to these primary findings, our work makes the following contributions to the field: **(4)** the first release of an instruction-tuned family of LLMs specifically for Basque, in sizes of 8B and 70B parameters, the latter of which proved competitive with GPT-4o and Claude Sonnet in the arena; **(5)** the release of large-scale, synthetic instruction-tuning datasets in English and Basque; and **(6)** the release of the first preference dataset in Basque, containing real user prompts, model responses, and preference annotations that could support future preference alignment research. Through these contributions, we aim to advance the state of language technology for Basque while establishing methodologies applicable to other low-resource languages.

2 Related Work

Research on developing LLMs for under-resourced languages has explored various approaches, with varying degrees of success. Initial attempts to develop models from scratch for specific low-resource languages have proven challenging due to limited training data. Multilingual model development has emerged as a more promising strategy, with researchers leveraging cross-lingual transfer learning to improve performance (Scao et al., 2023; Le Scao et al., 2022; Shliazhko et al., 2024). The most effective approach to date involves continued pretraining of existing multilingual models, which allows for language-specific adaptation while benefiting from the rich linguistic representations of larger training corpora (Etxaniz et al., 2024b; Luukkonen et al., 2023; Tran et al., 2024). While progress has been made in developing these base models, optimal methods for instructing and fine-tuning them for under-resourced languages remain largely unexplored (Gonzalez-Agirre et al., 2025; Martins et al., 2025; Üstün et al., 2024).

Instruction-tuning for under-resourced languages has explored various approaches to overcome the scarcity of native instruction data. Different studies leverage either English-centric pretrained models or multilingual models as pivot architectures for cross-lingual transfer (Purason et al., 2025). Regarding the data, researchers have explored incorporating multilingual instruction datasets that include limited coverage of lower-resourced languages (Shaham et al., 2024); translating existing English instruction sets into target languages either automatically or with human verification (Joshi et al., 2025; Zosa et al., 2025); and applying data augmentation techniques like

[2]commoncrawl.github.io/cc-crawl-statistics/plots/languages.html

back-translation, language-specific prompting, and template-based instruction generation to expand limited resources (Li et al., 2024). Additionally, cross-lingual in-context learning has shown interesting results (Cahyawijaya et al., 2024).

Regarding Basque language adaptation, two significant studies have been conducted. Etxaniz et al. (2024b) developed Latxa by adapting Llama 2 models through continued pretraining. Meanwhile, Corral et al. (2025) created Llama-eus by adapting Llama 3.1 and subsequently performing both instruction tuning and preference alignment using machine-translated data, adhering to widely accepted methodologies. However, the former focuses solely on foundation models, without considering instruction-tuned models, while the latter implements only a single strategy for instructing an adapted language model. In contrast, this work explores multiple strategies and combinations systematically to effectively instruct (or adapt) a language model for Basque.

3 Resources

Instructing LLMs typically relies on two components: base (or foundational) LLMs and instruction datasets. For non-hegemonic languages, obtaining instruction datasets can be very challenging, particularly in low-resource language scenarios. In the case of Basque in particular, there are no manually generated, or even good quality automatically generated, large sets of instruction-answer pairs. Consequently, as shown in Fig. 1, our available resources are constrained to corpora on the target language, and base and instruct models for high-resource languages. From these limited resources, we derive the necessary components to create Basque instruction-tuned models through strategic combinations of synthetic data generation and model adaptation. In the following sections, we describe these seed resources and derivations.

3.1 Basque corpora

For the pretraining data, we have leveraged the corpora used to train Latxa, the first family of LLMs trained specifically for Basque (Etxaniz et al., 2024b). This corpus comprises 4.3M of high-quality documents in Basque, roughly 3.5B Llama 3.1 tokens. Among the sources, it contains high-quality news data extracted using ad-hoc scrapers (Artetxe et al., 2022), Wikipedia[3] and sources based on Common Crawl such as CulturaX (Nguyen et al., 2024), Colossal OSCAR (Abadji et al., 2022) and HLPT v1.1 (de Gibert et al., 2024). This corpus comes normalized, deduplicated and filtered. The data is publicly available in the HuggingFace hub.[4] We will henceforth refer to this corpus as $\text{Corpus}_{\text{EU}}$.

3.2 Backbone models

As our base LLM (i.e., models that have not been fine-tuned to follow chat-style instructions) we use Llama 3.1 (Grattafiori et al., 2024). Llama 3.1 is a publicly available model widely adopted by the community due to its strong performance across English and other high-resource languages. We refer to this model as BASE_{EN} throughout the paper. In addition, following Etxaniz et al. (2024b), we train a new Latxa model based on Llama 3.1, which we denote as BASE_{EU}. For the instruction-tuned models, we adopt a similar strategy and use the instruction-following version of Llama 3.1, which we refer to as $\text{INSTRUCT}_{\text{EN}}$.

3.3 Instruction Sampling and Translation

Existing (English) instruction datasets rely on either high-quality, manually crafted instructions and responses (e.g., No Robots),[5] fully automatically generated instructions and responses (Ding et al., 2023; Ge et al., 2025), or a combination of both, such as manually written prompts paired with automatically generated responses (Zhao et al., 2024). Using any of these datasets would introduce an additional confounding factor into our analysis (namely, knowledge distilled from a powerful LLM), which could lead a model trained on such data to outperform our $\text{INSTRUCT}_{\text{EN}}$, thus introducing noise into our evaluation. This would raise a separate research question that falls outside the scope of this paper: what is the best (combination of) instruction dataset(s) to train a model on? In the case of Basque, however, there is no publicly available set of instructions. The following paragraphs detail the process of generating the instructions for each language.

English instructions. To avoid external influences, we instead sample instructions directly from our $\text{INSTRUCT}_{\text{EN}}$ model. We generate the English instructions following (Xu et al., 2025). Using this technique, we conditioned $\text{INSTRUCT}_{\text{EN}}$ to

[3]The `20231101` dump corresponding to November 2023.

[4]hf.co/datasets/HiTZ/latxa-corpus-v1.1

[5]hf.co/datasets/HuggingFaceH4/no_robots

generate instructions of different types and tasks: *general-purpose*, *code*, *math*, *arithmetic* and *translation*. We generated a total of 4M English instructions. However, after a hyperparameter search, we found out that using just 1M instructions yielded better results overall (see Appendix B). We share more details and examples of the process in Appendix A.

Basque instructions. We translated instructions sampled from $\textsc{Instruct}_{\text{EN}}$ using few-shot prompting with $\textsc{Base}_{\text{EU}}$. Existing machine translation systems for the English–Basque language pair (e.g., NLLB (Costa-jussà et al., 2022)) are primarily trained on sentence-level textual data and often struggle with more complex inputs, including selectively translating natural language content embedded within code snippets. By leveraging an LLM like $\textsc{Base}_{\text{EU}}$, which has been exposed to diverse data types, we obtained higher-quality translations for this setting. Moreover, using a model trained within our own experimental framework allows us to avoid introducing external factors into our pipeline. More details about the process and prompts used to translate the instructions are given in Appendix A.

4 Experimental Setup

We formalize our experimental setup as follows. Let $\mathcal{M} = \{\textsc{Base}_{\text{EN}}, \textsc{Base}_{\text{EU}}, \textsc{Instruct}_{\text{EN}}\}$ be the set of backbone models and $\mathcal{D} = \{\text{Corpus}_{\text{EU}}, \text{Instructions}_{\text{EN}}, \text{Instructions}_{\text{EU}}\}$ be the set of binary variables indicating whether to use Basque corpora, English instructions, and/or Basque instructions. The space of possible configurations is thus $\mathcal{M} \times \mathcal{P}(\mathcal{D})$, where $\mathcal{P}(\mathcal{D})$ is the power set of $\mathcal{D}$, yielding $|\mathcal{M}| \times 2^{|\mathcal{D}|} = 3 \times 2^3 = 24$ theoretical combinations. Note that we explore training strategies that leverage both **raw text and instruction data simultaneously**. From the total of 24 combinations, we exclude redundant configurations where a model is retrained on data it was originally trained with. The resulting set of distinct instruction-tuned model variants therefore comprises 18 configurations: the original Llama 3.1 Instruct 8B (i.e., $\textsc{Instruct}_{\text{EN}}$) and 17 new 8B-sized models. Table 3 in Appendix B provides a complete account of all model variants and their shorthand names. Additionally, we trained a 70B model following the configuration that performed best in preliminary benchmark evaluations.

Regarding the baselines, the primary baseline in our analysis is the $\textsc{Instruct}_{\text{EN}}$ model, as it is the only backbone capable of following instructions. However, since we examine the effect of each variable in $\mathcal{D}$ individually, the specific points of comparison used vary across cases. For additional context, we also evaluate two proprietary models known for their strong performance in Basque:[6] OpenAI's GPT-4o[7] and Anthropic's Claude 3.5 Sonnet.[8]

5 Evaluation

We employed two complementary evaluation approaches to assess the impact of each instruction-tuning recipe. On the one hand, we used a selection of static benchmarks that evaluate specific model capabilities and knowledge through standardized tests. On the other hand, we conducted human evaluations through A/B testing (arena style) to capture qualitative aspects of model performance. In addition, we look into the impact of our recipes on safety and bias.

5.1 Static Benchmarks

We selected benchmarks that are close to real use cases, from a varied range of categories: *reading comprehension*, *common sense*, *linguistic proficiency*, *knowledge* and *maths & reasoning*.

For each benchmark, where possible, we evaluated the Basque, English, and Spanish versions to facilitate the analysis of language-specific trade-offs for each fine-tuned model variant. This choice of evaluation languages reflects the linguistic reality of the Basque-speaking community in northern Spain, where Basque and Spanish are co-official and English is the most commonly taught foreign language. Importantly, these languages come from distinct families: Basque is a language isolate, Spanish is Romance, and English is Germanic. Thus, we examine cross-lingual transfer effects and assess whether improvements in our language come at the cost of performance in related community languages, including one—Spanish—not directly targeted by our experiments. In total, then, we evaluated 27 benchmarks, as detailed in Appendix C.1.

For conducting these evaluations, we relied on LM Evaluation Harness (Biderman et al., 2024).

[6] We did not include the instructed Llama-eus (Corral et al., 2025) in our evaluation, as it was not publicly available at the time of experimentation.

[7] `gpt-4o-2024-11-20`

[8] `claude-3-5-sonnet-20241022`

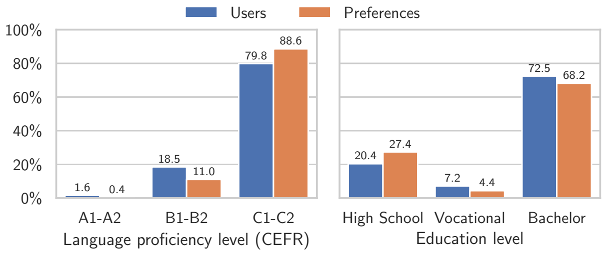


Figure 2: Distribution of participants and preferences by education level and language proficiency level.

Most datasets are framed as multiple-choice problems where models' answers are determined by selecting the option with the highest log probability. For generative tasks, answers are directly sampled from the model. To provide models with contextual examples, our evaluations employed a few-shot setting. All results are measured for accuracy following standard, public implementations. Refer to Appendix C.1 and our repository for details.

When evaluating proprietary models, we cannot directly compute log probabilities because we have no access to model weights. This limitation restricts our evaluation to only those benchmarks implemented as explicit letter-choice questions (A, B, C, ...) and the free-form generative task MGSM, excluding benchmarks that require comparing verbalized option likelihoods. For the compatible multiple-choice benchmarks, we prompt models to output a single letter as their answer, using the same prompts and few-shot examples as with open models to maintain comparability.

5.2 Human Evaluation: Arena

Unlike static benchmarks, which rely on fixed datasets and automatic metrics, arena-style evaluations are better suited for assessing open-ended text generation, where subjective quality judgments play a central role. In this section, we first describe our implementation of the arena framework, including details on participants and evaluation conditions. For additional details about the human evaluation, refer to Appendix C.2, where we describe our infrastructure and introduce the Bradley–Terry model (Bradley and Terry, 1952), which we use to infer a model ranking from the collected pairwise preferences.

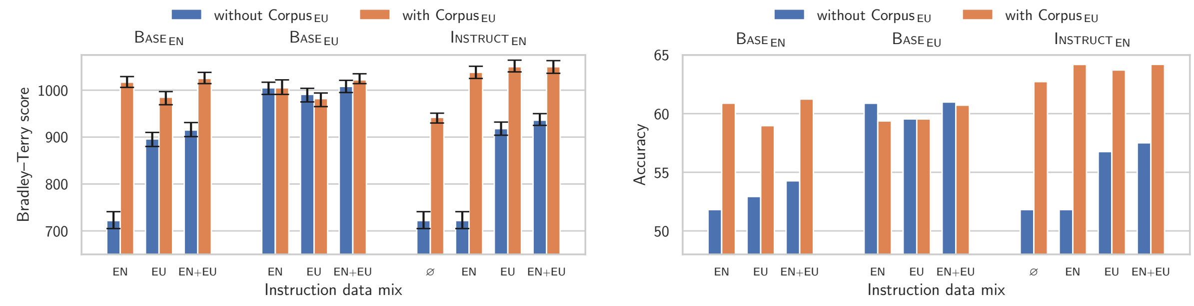
To gather human preferences for our evaluation, we organized a community-driven initiative. This collaborative effort ran for 14 days and attracted approximately 1,680 participants, resulting in a total of 12,890 preference annotations. The event was open to any Basque speaker, regardless of their proficiency level. Participants were required to register their educational background and language proficiency before contributing. Once registered, users could submit prompts and compare model responses. Fig. 2 shows that the majority of the participants—and, therefore, the preferences—have a bachelor or superior education and a high or native language proficiency level.

In the annotation process, participants evaluated pairs of model responses by making a three-way choice (i.e., prefer model 'A', prefer model 'B', or consider them tied) across two dimensions: content quality and linguistic quality. Linguistic quality was considered as a separate dimension because not all models produce fluent and sound Basque—an uncommon issue among high-resource languages. In cases where participants' judgments were contradictory between the dimensions, a third question about overall quality was presented to determine the final choice.

It merits mention that we offered prizes based on user activity to encourage widespread participation and maximize the number and diversity of collected preferences. While this strategy succeeded in increasing engagement, it also attracted malicious or dishonest users who prioritized quantity over quality to win rewards. We proactively identified and banned such actors using a combination of heuristics and manual review.

5.3 Safety and Bias

This paper does not address the alignment step typically required to reduce biases and prevent unsafe responses in production-ready LLMs. However, the backbone model INSTRUCT$_{\text{EN}}$ is already aligned, and we expect variants based on this backbone to retain safety-related behaviors. To verify this, we constructed a new Basque-English parallel dataset inspired by XSTest (Röttger et al., 2024), combining clearly unsafe prompts with superficially similar but safe ones. We measure both Violation rates (VR) and False Refusal rates (FRR), where the model wrongly declines safe prompts. For bias, we rely on BBQ (Parrish et al., 2022) and its adaptation to Basque (Zulaika and Saralegi, 2025), reporting results in terms of accuracy. Further details are available in Appendix C.3.



(a) Bradley–Terry scores in the human evaluation arena

(b) Average accuracy results in Basque benchmarks

Figure 3: Performance comparison of instruction-tuned models across different dimensions: backbone model, Corpus$_{\text{EU}}$ usage and instruction data composition. Error bars in Fig. 3a indicate 90% confidence intervals.

6 Results

In the following paragraphs, we disclose the effect of each component: backbone models, Basque corpora and bilingual instructions. For both the benchmark and the human evaluation, main results are presented in Fig. 3. Complete results for all tested models, broken down by benchmark and language, along with detailed arena evaluation results, can be consulted in Appendix D.

6.1 The Impact of Basque Corpora

We begin by analyzing the influence of Corpus$_{\text{EU}}$, as it is intuitively the most critical resource for teaching a new language to an LLM. Fig. 3 confirms that this intuition aligns well with empirical results. In both human evaluations (Fig. 3a) and benchmark scores (Fig. 3b), models trained on Corpus$_{\text{EU}}$ achieved significantly better performance. The advantage is especially pronounced—up to 12 points in accuracy and over 300 points in arena score—when no other Basque signal (i.e., Basque instructions) is included. However, for models that already use BASE$_{\text{EU}}$ as their backbone, additional exposure to Corpus$_{\text{EU}}$ offered little benefit and was sometimes even detrimental.

We conclude that **using target language corpora is highly beneficial and possibly essential for training an instruction-tuned LLM in our low-resource language**. Therefore, the following analyses will focus exclusively on the variants trained with Corpus$_{\text{EU}}$.

6.2 The Impact of Instruction Data

We analyze the effect of instruction data by comparing variants trained with no instructions (∅), English-only instructions (I$_{\text{EU}}$), Basque-only instructions (I$_{\text{EN}}$), and their combination (I$_{\text{EN+EU}}$).

Starting with the question of whether to use instructions at all, we focus on INSTRUCT$_{\text{EN}}$-based variants. Human evaluation results (Fig. 3a) clearly show that incorporating instructions, regardless of language, helps mitigate catastrophic forgetting and improves arena scores by nearly 100 points. In contrast, benchmark results (Fig. 3b) show only marginal gains, particularly when using Basque instructions. This discrepancy highlights the limitations of static benchmarks and underscores the value of human or text generation-based evaluation.

When comparing instruction languages, we observe a general trend: English instructions tend to yield better results across both evaluation methods. However, there are exceptions. For example, models based on INSTRUCT$_{\text{EN}}$ perform comparably or slightly better with Basque instructions in human evaluations, while BASE$_{\text{EU}}$-based models perform similarly on benchmarks.

Notably, combining English and Basque instructions consistently produces the best results across most scenarios. While this improvement could be attributed to certain model variants having access to more training data, our preliminary results in Appendix B refute this hypothesis, as using more monolingual instructions (1M vs 4M) resulted in similar results.

Although the improvements are not always significantly better, we conclude that **including instructions in both languages results in more robust models**, achieving stronger performance regardless of the backbone. Consequently, the remainder of our analysis will focus on models trained with bilingual instruction data.

	8B				70B		Proprietary	
	INSTRUCT EN	+ C EU I EN	+ C EU I EU	+ C EU I EN+EU	INSTRUCT EN	+ C EU I EN	3.5 Sonnet	GPT-4o
Belebele	73.89	80.00	81.44	**83.00**	89.11	**91.00**	**94.22**	92.88
BertaQA Global	67.10	**74.62**	73.54	72.99	83.53	**87.42**	**93.52**	91.01
BertaQA Local	44.97	65.23	**66.07**	65.57	53.51	**77.71**	**80.45**	74.83
EusProficiency	34.13	**52.83**	52.06	52.35	43.59	**68.00**	**81.60**	74.25
EusReading	49.72	59.66	**62.78**	61.93	72.16	**78.98**	**87.39**	84.38
EusTrivia	45.01	61.05	**62.33**	62.10	62.51	**74.17**	**84.60**	80.70
EusExams	46.21	56.00	56.01	**56.23**	63.28	**71.56**	**82.68**	79.17
MGSM	45.60	**54.00**	46.40	50.80	76.40	**80.00**	**85.20**	79.20
MMLU	50.37	**57.04**	52.96	56.30	68.52	**68.89**	**79.63**	76.66
Benchmark Avg	50.78	62.27	61.51	**62.36**	68.07	**77.53**	**85.48**	81.45
Arena Content	766 (-17,+14)	1031 (-12,+15)	1045 (-13,+11)	1047 (-12,+12)	-	1127 (-11,+10)	1150 (-17,+12)	1183 (-13,+15)
Arena Language	783 (-12,+12)	1036 (-10,+11)	1034 (-10, +8)	1038 (-8,+10)	-	1083 (-13,+13)	1082 (-11,+11)	1093 (-10,+12)
Arena Global	722 (-17,+19)	1038 (-13,+13)	1050 (-11,+14)	1050 (-14,+13)	-	1141 (-11,+15)	1153 (-21,+13)	1188 (-17,+13)

Table 1: Results for the baseline (INSTRUCT EN), the best 3 performing variants, 70B models and proprietary models. Best results among comparable setups are marked in **bold**. Arena scores are given with 90% confidence intervals.

6.3 The Impact of Backbone Models and Curriculum Learning

By analyzing models trained from different backbones, we explore various curriculum learning strategies: (i) teaching the language first and then instruction following, (ii) teaching instruction following in English first and then the target language, or (iii) learning everything simultaneously.

Language first vs. simultaneously. When comparing models based on BASE EN (i.e., acquiring Basque and instruction-following capabilities simultaneously) with those based on BASE EU (learning the language first, then instruction following), we observe no significant difference in performance. Interestingly, the BASE EU variant *without* access to Corpus EU during instruction tuning achieves performance nearly identical to the BASE EN variant *with* access to Corpus EU, across all instruction settings. This suggests that teaching the language in a separate pretraining step offers no measurable advantage. From this, we conclude that **there is no compelling reason to separate language acquisition from instruction tuning**.

Instructions first vs. simultaneously. While instruction tuning pipelines are often complex and multi-staged (Lambert et al., 2025), our approach adopts a simpler structure. Previous work by Xu et al. (2025) showed that models initialized from BASE EN and trained with sampled instructions from INSTRUCT EN perform comparably to INSTRUCT EN. However, our findings indicate that this strategy does *not* transfer well to low-resource languages. As shown in Fig. 3, models based on INSTRUCT EN consistently outperform those based on BASE EN, both in human evaluations and benchmarks. These results support the conclusion that **starting from a well-instructed English backbone yields better performance than learning everything from scratch**.

6.4 The Impact of the Scale

Based on our previous analyses, we scaled up the variant that performed best in preliminary benchmark evaluations.[9] Table 1 shows the results for some multiple-choice benchmarks and arena scores. On the one hand, we have the three best-performing 8B variants and the baseline. On the other hand, we present the results for the 70B best-performing variant and the baseline. Finally, we also show the performance of Claude 3.5 Sonnet and GPT-4o.

Scaling to bigger sizes. We analyzed the effect of our language adaptation process when training a larger model. Despite the results of the 70B INSTRUCT EN baseline being significantly better than the 8B counterpart (even surpassing the INSTRUCT EU variants), we observe that our language adaptation step still had similar improvements to those obtained with 8B sized models—almost 10 accuracy points gain on average. The biggest gains are observed in local knowledge and language proficiency, the only benchmarks where the 70B INSTRUCT EN underperforms the best 8B variant.

Comparing to the State of the Art. When compared to the leading commercial models in Basque, our best model falls slightly behind in most benchmarks except for BertaQA Local and MGSM,

[9]We did this before running the arena as we wanted to include a 70B model in the human evaluation.

where our model performs better than GPT-4o—particularly in local knowledge about the Basque Country. Regarding the arena score, **our best model is on par with the commercial models in perceived linguistic quality**, but falls behind the best model in the content quality and global scores. Interestingly, Claude 3.5 Sonnet outperforms GPT-4o in all benchmarks, but the latter gets a higher score in the arena. Our best model being worse than SotA despite focusing on Basque might be related to the weaker backbone model we used, as INSTRUCT $_{\mathrm{EN}}$ is overall a weaker model. Using a stronger and larger backbone model in the future could lead to results that match the SotA models in benchmarks and arena score.

7 Analysis and Discussion

In this section, we provide additional analysis and discussion of our results. First, we focus on the correlation of our two main evaluation methods. Then, we measure the trade-off between Basque and other languages. Finally, we analyze the safety and biases of our models.

Benchmark–Arena correlation. Fig. 4 shows Spearman's rank correlation coefficients, ρ, between benchmark performance and arena scores, across different benchmark languages (Basque, English, Spanish) and arena dimensions (content, language, and global). We observe that average Basque benchmark performance and arena rankings correlate strongly, with $\rho > 0.80$ across all arena dimensions—which suggests that automated Basque benchmarks may provide a reliable proxy for human evaluation in future research. This correlation is particularly pronounced for specific benchmarks, including EusProficiency, EusTrivia, EusExams and BertaQA, which are interestingly the datasets that were natively constructed in Basque, rather than translated from existing English benchmarks. The average of English benchmarks displays overall positive but non-significant correlations. Only BertaQA shows positive correlations, with the local subset obtaining correlations similar to the Basque BertaQA, likely reflecting the types of culturally-specific questions that users posed in the arena. Spanish benchmarks show, on average, no correlation with the arena. Some of the English and Spanish benchmarks show large negative correlations, reflecting the performance trade-off between Basque and other languages.

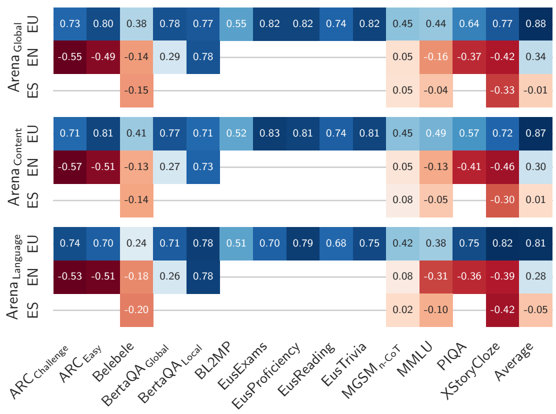


Figure 4: Spearman's rank correlation coefficients between benchmark performance and arena evaluation dimensions.

Trade-off between Basque and other languages. Fig. 5 shows performance changes across languages relative to each backbone model on the multilingual benchmarks Belebele, MGSM, MMLU, and XStoryCloze. We observe that INSTRUCT $_{\mathrm{EN}}$-based models exhibit a clear trade-off: improvements in Basque come at a cost of decreased performance in English and Spanish, suggesting a competitive relationship between languages in the model's parameter space. In contrast, greater flexibility for multilingual adaptation is observed in BASE $_{\mathrm{EN}}$ models, which improve across all three languages (though from a lower absolute performance baseline). BASE $_{\mathrm{EU}}$ models show moderate changes with gains primarily in Spanish and English rather than Basque. As observed previously, Corpus $_{\mathrm{EU}}$ consistently yields the largest performance gains for the target language. Among INSTRUCT $_{\mathrm{EN}}$ variants, the configuration with Corpus $_{\mathrm{EU}}$ and Instructions $_{\mathrm{EN}}$ achieves the most Basque improvement and the least regression in other languages. Despite these adaptation strategies, models still perform better in English and Spanish on equivalent tasks, with the exception of culturally-specific knowledge as evidenced by the results on the BertaQA dataset (see complete results in Appendix D.1).

Safety and bias. We evaluate the model variant INS $_{\mathrm{EN}}$ C $_{\mathrm{EU}}$ I $_{\mathrm{EN}}$, comparing it with two critical counterparts: (1) BAS $_{\mathrm{EN}}$ C $_{\mathrm{EU}}$ I $_{\mathrm{EN}}$, to analyze the impact of starting from an already instruction-tuned backbone versus a base model (8B only); and (2) INSTRUCT $_{\mathrm{EN}}$, to measure potential alignment changes introduced with our

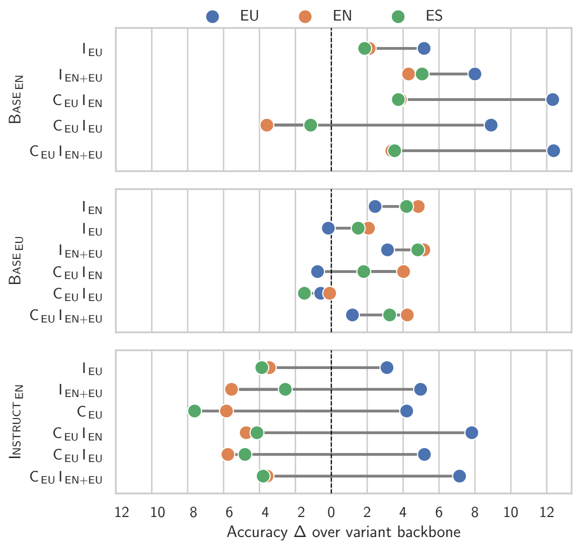


Figure 5: Accuracy differences for model variants compared to their respective backbones across Belebele, MGSM, MMLU, and XStoryCloze averages by language. Positive values indicate performance improvements; negative values indicate regression.

fine-tuning data mix (8B and 70B). As shown in Table 2, the BAS$_{EN}$ C$_{EU}$ I$_{EN}$ model demonstrates low safety and high bias in the Basque language compared to English, reflecting a significant lack of alignment. In contrast, INSTRUCT$_{EN}$ and its variant INS$_{EN}$ C$_{EU}$ I$_{EN}$ are better aligned in both languages. Interestingly, in Basque, the 8B INSTRUCT$_{EN}$ shows a lower VR than both BAS$_{EN}$ C$_{EU}$ I$_{EN}$ and INS$_{EN}$ C$_{EU}$ I$_{EN}$ models and even the larger 70B counterpart. This behavior is due to the model's limited comprehension of instructions in the Basque language, leading it to produce inconsistent yet safe responses. On the other hand, the 70B models display comparable performance across both languages, with the INS$_{EN}$ C$_{EU}$ I$_{EN}$ version slightly outperforming others and achieving a lower VR in Basque. Notably, the FRR remains very low across all models, indicating that safety mechanisms do not come at the cost of excessive conservatism. Bias outcomes in English are consistently better than in Basque, and larger models generally perform better than smaller ones. However, the differences in bias between the original and language-adapted models are minimal. Overall, we demonstrate that much of the safety and bias alignment is transferred from INSTRUCT$_{EN}$, not only in the newly added language, but also in the predominant language.

		EU			EN		
		VR ↓	FRR ↓	BBQ ↑	VR ↓	FRR ↓	BBQ ↑
8B	**INS$_{EN}$ C$_{EU}$ I$_{EN}$**	24.00	0.00	70.80	16.00	0.00	87.06
	BAS$_{EN}$ C$_{EU}$ I$_{EN}$	44.00	0.00	51.73	20.00	4.00	72.35
	INSTRUCT$_{EN}$	8.00	0.00	71.03	4.00	0.00	87.65
70B	**INS$_{EN}$ C$_{EU}$ I$_{EN}$**	4.00	0.00	85.29	4.00	0.00	94.38
	INSTRUCT$_{EN}$	20.00	0.00	84.90	8.00	0.00	95.32

Table 2: Safety and bias results for Basque and English datasets. Safety is measured in terms of violation rate (VR) and false refusal rate (FRR) where ↓ indicates lower values are better. Bias is measured with BBQ accuracy, where ↑ indicates higher values are better.

8 Conclusions

This systematic study on instruction-tuning LLMs for Basque reveals several key strategies for low-resource language adaptation. We found that **target language corpora are essential** for effective learning. Employing **bilingual (English and Basque) synthetic instructions yielded the most robust models** whereas English-only instructions remain competitive. Crucially, **starting from an instruction-tuned English model and adapting it to Basque proved more effective** than training a base model for both language and instruction-following, or pretraining for language separately before instruction tuning. Our work contributes new Basque models, open instruction and human preference datasets, and methodological insights to guide future low-resource LLM development.

In the future, we plan to extend the exploration using instructions created by humans. The preference data we release can also be used to align the models. Scaling to stronger backbones will also lead to better results that could match the SotA commercial models in Basque. The strong correlation observed between aggregated Basque benchmarks and human evaluations also suggests a path for more efficient proxy evaluations.

Limitations

This study presents a systematic analysis aimed at identifying the most effective method for developing an instruction-tuned model for a low-resource language. However, due to the combinatorial nature of such analyses, we had to constrain certain dimensions of our exploration, as adding any additional axis would effectively double the amount of work required, including costly human evaluations.

Our first limitation is the choice of language. We focused on Basque, primarily because it is a low-resource language with just enough available data to enable language adaptation of base models (Etxaniz et al., 2024b). While some other languages have significantly fewer resources than Basque, our conclusions may not fully generalize to those more extreme low-resource scenarios.

The second limitation is the choice of model family. We conducted all experiments using the Llama 3.1 family as the backbone. This decision was motivated by its widespread adoption and its existing ability to generate text in Basque, although often with substantial linguistic errors. Evaluating more recent or higher-performing models could slightly influence our findings. However, to the best of our knowledge, there is currently no open model family capable of producing linguistically correct Basque.

Third, this study primarily focuses on the initial instruction-tuning phase. While we did collect preference data, we did not extend our analysis to include preference alignment techniques. Including this additional phase would again have doubled the experimental workload and human evaluation requirements.

Finally, although we did not perform analyses on potential data contamination issues, previous work on which our work is based took measures against contamination (Etxaniz et al., 2024a,b).

Acknowledgements

This work has been partially supported by the Basque Government (Research group funding IT1570-22 and IKER-GAITU project), the Spanish Ministry for Digital Transformation and of Civil Service, and the EU-funded NextGenerationEU Recovery, Transformation and Resilience Plan (ILENIA project, 2022/TL22/00215335; and ALIA project). The models were trained on the Leonardo supercomputer at CINECA under the EuroHPC Joint Undertaking, project EHPC-EXT-2024E01-042. Julen Etxaniz and Ekhi Azurmendi hold a PhD grant from the Basque Government (PRE_2024_2_0028 and PRE_2024_1_0035, respectively).

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.

Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*.

Mikel Artetxe, Itziar Aldabe, Rodrigo Agerri, Olatz Perez-de Viñaspre, and Aitor Soroa. 2022. Does corpus quality really matter for low-resource languages? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7383–7390, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Irene Baucells, Javier Aula-Blasco, Iria de Dios-Flores, Silvia Paniagua Suárez, Naiara Perez, Anna Salles, Susana Sotelo Docio, Júlia Falcão, Jose Javier Saiz, Robiert Sepulveda Torres, Jeremy Barnes, Pablo Gamallo, Aitor Gonzalez-Agirre, German Rigau, and Marta Villegas. 2025. IberoBench: A benchmark for LLM evaluation in Iberian languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10491–10519, Abu Dhabi, UAE. Association for Computational Linguistics.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, and 11 others. 2024. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*.

Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. LLMs are few-shot in-context low-resource language learners. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 405–433, Mexico City, Mexico. Association for Computational Linguistics.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 8359–8388. PMLR.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Ander Corral, Ixak Sarasua Antero, and Xabier Saralegi. 2025. Pipeline analysis for developing instruct LLMs in low-resource languages: A case study on Basque. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12636–12655, Albuquerque, New Mexico. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 19 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, Sampo Pyysalo, Stephan Oepen, and Jörg Tiedemann. 2024. A new massive multilingual dataset for high-performance language technologies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1116–1128, Torino, Italia. ELRA and ICCL.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051, Singapore. Association for Computational Linguistics.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2024a. Bertaqa: How much do language models know about local culture? *NeurIPS Datasets and Benchmarks 2024*, 37:34077–34097.

Julen Etxaniz, Oscar Sainz, Naiara Miguel, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, and Aitor Soroa. 2024b. Latxa: An open language model and evaluation suite for Basque. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14952–14972, Bangkok, Thailand. Association for Computational Linguistics.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2025. Scaling synthetic data creation with 1,000,000,000 personas. *Preprint*, arXiv:2406.20094.

Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, Mario Mina, Adrián Rubio, Alexander Shvets, Anna Sallés, Iñaki Lacunza, Iñigo Pikabea, Jorge Palomar, Júlia Falcão, Lucía Tormo, and 4 others. 2025. Salamandra technical report. *Preprint*, arXiv:2502.08489.

María Grandury, Javier Aula-Blasco, Júlia Falcão, Clémentine Fourrier, Miguel González, Gonzalo Martínez, Gonzalo Santamaría, Rodrigo Agerri, Nuria Aldama, Luis Chiruzzo, Javier Conde, Helena Gómez, Marta Guerrero, Guido Ivetta11, Natalia López, Flor Miriam Plaza del Arco, María Teresa Martín-Valdivia, Helena Montoro, Carmen Muñoz, and 5 others. 2025. La leaderboard: A large language model leaderboard for spanish varieties and languages of spain and latin america. In *NAACL 2025 Workshop on Language Models for Underserved Communities*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Eric Harper, Somshubra Majumdar, Oleksii Kuchaiev, Li Jason, Yang Zhang, Evelina Bakhturina, Vahid Noroozi, Sandeep Subramanian, Koluguri Nithin, Huang Jocelyn, Fei Jia, Jagadeesh Balam, Xuesong Yang, Micha Livne, Yi Dong, Sean Naren, and Boris Ginsburg. 2024. Nemo: a toolkit for conversational ai and large language models.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Raviraj Joshi, Kanishk Singla, Anusha Kamath, Raunak Kalani, Rakesh Paul, Utkarsh Vaidya, Sanjay Singh

Chauhan, Niranjan Wartikar, and Eileen Long. 2025. Adapting multilingual LLMs to low-resource languages using continued pre-training and synthetic corpus: A case study for Hindi LLMs. In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 50–57, Abu Dhabi. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2025. Tülu 3: Pushing frontiers in open language model post-training. *Preprint*, arXiv:2411.15124.

Teven Le Scao, Thomas Wang, Daniel Hesslow, Stas Bekman, M Saiful Bari, Stella Biderman, Hady Elsahar, Niklas Muennighoff, Jason Phang, Ofir Press, Colin Raffel, Victor Sanh, Sheng Shen, Lintang Sutawika, Jaesung Tae, Zheng Xin Yong, Julien Launay, and Iz Beltagy. 2022. What language model to train if you have one million GPU hours? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 765–782, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chong Li, Wen Yang, Jiajun Zhang, Jinliang Lu, Shaonan Wang, and Chengqing Zong. 2024. X-instruction: Aligning language model in low-resource languages with self-curated cross-lingual instructions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 546–566, Bangkok, Thailand. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, and 2 others. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Risto Luukkonen, Ville Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kupari, Filip Ginter, Veronika Laippala, Niklas Muennighoff, Aleksandra Piktus, Thomas Wang, Nouamane Tazi, Teven Scao, Thomas Wolf, Osma Suominen, Samuli Sairanen, Mikko Merioksa, Jyrki Heinonen, Aija Vahtola, and 2 others. 2023. FinGPT: Large generative models for a small language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2710–2726, Singapore. Association for Computational Linguistics.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G.C. de Souza, Alexandra Birch, and André F.T. Martins. 2025. Eurollm: Multilingual language models for europe. *Procedia Computer Science*, 255:53–62. Proceedings of the Second EuroHPC user day.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2017. LSDSem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51, Valencia, Spain. Association for Computational Linguistics.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Taido Purason, Hele-Andra Kuulmets, and Mark Fishel. 2025. LLMs for extremely low-resource Finno-Ugric languages. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6677–6697, Albuquerque, New Mexico. Association for Computational Linguistics.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400, Mexico City, Mexico. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, and 373 others. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2304–2317, Bangkok, Thailand. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mGPT: Few-shot learners go multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79.

Khanh-Tung Tran, Barry O'Sullivan, and Hoang D. Nguyen. 2024. Uccix: Irish-excellence large language model. In *27th European Conference on Artificial Intelligence, 19–24 October 2024, Santiago de Compostela, Spain*.

Gorka Urbizu, Muitze Zulaika, Xabier Saralegi, and Ander Corral. 2024. How well can BERT learn the grammar of an agglutinative and flexible-order language? the case of Basque. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8334–8348, Torino, Italia. ELRA and ICCL.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2025. Magpie: Alignment data synthesis from scratch by prompting aligned LLMs with nothing. In *The Thirteenth International Conference on Learning Representations*.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. 2023. Pytorch fsdp: Experiences on scaling fully sharded data parallel. *Proc. VLDB Endow.*, 16(12):3848–3860.

Elaine Zosa, Ville Komulainen, and Sampo Pyysalo. 2025. Got compute, but no data: Lessons from post-training a Finnish LLM. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 826–832, Tallinn, Estonia. University of Tartu Library.

Muitze Zulaika and Xabier Saralegi. 2025. BasqBBQ: A QA benchmark for assessing social biases in LLMs for Basque, a low-resource language. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4753–4767, Abu Dhabi, UAE. Association for Computational Linguistics.

```
<|begin_of_text|>
<|start_header_id|>system<|end_header_id|>
Cutting Knowledge Date: December 2023
Today Date: 26 Jul 2024
<|eot_id|>
<|start_header_id|>user<|end_header_id|>
```

(a) General

```
<|begin_of_text|>
<|start_header_id|>system<|end_header_id|>
You are an AI assistant designed to provide
helpful, step-by-step guidance on solving
math problems. The user will ask you a wide
range of complex mathematical questions.
Your purpose is to assist users in
understanding mathematical concepts, working
through equations, and arriving at the
correct solutions.
<|eot_id|>
<|start_header_id|>user<|end_header_id|>
```

(b) Maths

```
<|begin_of_text|>
<|start_header_id|>system<|end_header_id|>
You are an AI assistant designed to provide
helpful, step-by-step guidance on solving
complex arithmetic operations. The user will
provide you with an arithmetic operation or
a concatenation of multiple arithmetic
operations. Your purpose is to assist users
in computing the results of the arithmetic
operation exlaining the process step by step.
<|eot_id|>
<|start_header_id|>user<|end_header_id|>
```

(c) Arithmetic

```
<|begin_of_text|>
<|start_header_id|>system<|end_header_id|>
You are an AI assistant designed to provide
helpful, step-by-step guidance on coding
problems. The user will ask you a wide range
of coding questions. Your purpose is to assist
users in understanding coding concepts,
working through code, and arriving at the
correct solutions.
<|eot_id|>
<|start_header_id|>user<|end_header_id|>
```

(d) Code

```
<|begin_of_text|>
<|start_header_id|>system<|end_header_id|>
You are an AI assistant specifically designed
to provide accurate and contextually
appropriate translations. Users will ask you
to translate a large text between various
languages. Your purpose is to translate the
text, maintaining the original context and
nuances.
<|eot_id|>
<|start_header_id|>user<|end_header_id|>
```

(e) Translation

Figure 6: Prompts used to generate the synthetic instructions

A Synthetic Instructions generation

A.1 English instructions

To generate the English synthetic instructions, we followed the Magpie technique (Xu et al., 2025). Briefly, it consists in letting the model generate text starting from the user's prompt instead of the assistant's response. See, for instance, Fig. 6a, where the model is asked to continue with the chat template immediately after the user header. We defined 5 prompts to generate different kinds of instructions (Fig. 6), then sampled instructions from the model using 10 different temperature values ranging 0.8–1.2. After generating the instructions, we applied the following filters:

1. **Duplicates:** keep unique instances.
2. **Repetitive prompts or responses:** remove instances with a sequence of tokens repeated over 100 times.
3. **Poor quality prompts or responses:** remove instances regarded as poor-quality by the model itself.
4. **Unfinished instructions:** we noticed that some instructions ended with ":", meaning that the instructions was incomplete. We removed those as well.

As we experimented with models of two sizes and did not want to have any external influence in our experiments, we generated the instructions twice, once per model size. We later used the instructions generated from the 8B model to train the 8B models, and those from 70B to train the 70B model. The total size in tokens of the English instructions used for training is 350M tokens. We are

```
You are a helpful AI assistant that specializes in English to Basque translations.
Your task is to translate instruction datasets from English to Basque.

Here are some important guidelines:
1. Maintain the original meaning and intent of the instructions.
2. Use standard Basque language (batua).
3. Keep the technical terms that don't have widely accepted Basque translations.
4. Preserve any code snippets, variables, or special characters exactly as they appear.
5. Translate only the text content, not the JSON structure.

The input will be a JSON object with English text. Please provide accurate Basque translations
for all text fields.

{% for example in fs_examples %}
English:
{{ example['english'] }}

Basque:
{{ example['basque'] }}
{% endfor %}

English:
{{ conversation }}

Basque:
```

Figure 7: Prompt used to translate the English instructions to Basque instructions.

releasing[10] the generated instructions (filtered and unfiltered) for replicability purposes.

A.2 Basque (translated) instructions

As discussed in Section 3.3, we generated Basque instructions by translating the English instructions using BASE$_{\text{EU}}$. The prompt we used to perform the translation is shown in Fig. 7. As for the few-shot examples, we included 2 conversations of *general* instructions and 1 conversation for each *math*, *arithmetic*, *code* and *translation* instruction types. These examples were sampled from the English instructions and post-edited manually. The exact examples are released in the GitHub repository.

Using a base LLM introduces several challenges when generating structured data. To mitigate issues with malformed outputs, we constrained the LLM's generation to adhere to a conversational JSON schema. Additionally, a non-negligible portion of the outputs (approximately 1%) exhibited repetition, either in the question or the response. To prevent these repetitive examples from being incorporated into the instruction-tuned model, we applied simple heuristics based on span frequency to identify and filter them out. The total size in tokens of the Basque instructions used for training is 503M tokens. We are releasing the translated instructions for replicability purposes.

A.3 Examples

Figs. 8 and 9 show a instruction and response pair generated and translated by our models in English and Basque respectively.

A.4 Manual evaluation

We conducted a manual analysis on a small sample of instructions to validate the translation pipeline. Two key insights emerged from this analysis:

- The quality of the original English instructions is not always optimal, particularly for translation-related tasks.
- The translated instructions are generally of high quality, often significantly better than NLLB translations for document-level, code-related, and math-related tasks. However, translations sometimes modify the intent slightly. For instance, the English instruction "Translate this sentence to English: [...]" was translated into Basque as *"Itzuli esaldi hau euskarara: [...]"*, which literally means "Translate this sentence to Basque: [...]". Nevertheless, the instruction remains well-adapted to the task as a whole and does not degrade overall quality.

[10] hf.co/collections/HiTZ/latxa-instruct-682f356091452b0028380804

```
User:
What's the difference between primary vs secondary market research?

Assistant:
Primary market research involves collecting original data directly from the source, such as
through surveys, focus groups, or experiments. This type of research is typically more expensive
and time-consuming but provides unique and specific insights.

Secondary market research, on the other hand, involves analyzing existing data that has already
been collected by someone else, such as academic studies, industry reports, or government
statistics. This type of research is often less expensive and faster to conduct but may not
provide the same level of specificity or relevance as primary research.

Here are some key differences between primary and secondary market research:

**Primary Market Research**

* Collects original data directly from the source
* Can be tailored to specific research questions or objectives
* Provides unique and specific insights
* Typically more expensive and time-consuming
* Examples: surveys, focus groups, experiments, customer interviews

**Secondary Market Research**

* Analyzes existing data collected by someone else
* Often less expensive and faster to conduct
* May not provide the same level of specificity or relevance
* Examples: academic studies, industry reports, government statistics, online reviews,
social media analytics

When deciding between primary and secondary market research, consider the following factors:

1. **Research objectives**: If you need specific, detailed information about your target market,
primary research may be the better choice. For general information or background research,
secondary research may suffice.
2. **Budget**: Primary research can be costly, while secondary research is often less expensive.
3. **Time constraints**: Secondary research is generally faster to conduct than primary research.
4. **Data quality**: Primary research provides original data, while secondary research relies on
existing data that may be outdated or biased.

Ultimately, a combination of both primary and secondary market research can provide a
comprehensive understanding of your target market.
```

Figure 8: English instruction generated by the 70B INSTRUCT$_{\text{EN}}$ model

In general terms, however, we concluded that reliably estimating instruction quality is non-trivial. While we are actively exploring this direction, it remains outside the current scope of the paper.

B Training details

Hardware and carbon footprint. Table 4 summarizes the training costs of our experiments (see Table 3 for the variants nomenclature). In total, we trained 19 models: one BASE$_{\text{EU}}$, seventeen INSTRUCT$_{\text{EU}}$ variants, and one 70B INSTRUCT$_{\text{EU}}$. Due to unforeseen circumstances, the BASE$_{\text{EU}}$ and INSTRUCT$_{\text{EU}}$ models were trained using different frameworks and infrastructure. The BASE$_{\text{EU}}$ model was trained with NeMo (Harper et al., 2024) on 64GB H100 GPUs provided by MareNostrum 5. In contrast, all INSTRUCT$_{\text{EU}}$ variants were trained using Fully Sharded Data-Parallel (Zhao et al., 2023) on 64GB A100 GPUs provided by CINECA Leonardo. The total compute time across all experiments amounted to $39,879.1$ GPU hours, which corresponds to $4,729.11$ kg CO_2eq, based on carbon intensity estimates from ElectricityMaps.[11]

Hyperparameters. Table 5 outlines the key hyperparameters used during training. Both 8B and

[11] At the time of the experiments: 0.297 kg/kWh for Italy and 0.157 kg/kWh otherwise, according to https://www.electricitymaps.com/.

Training Data	Backbone model		
	BASE$_{\text{EN}}$	**BASE$_{\text{EU}}$**	**INSTRUCT$_{\text{EN}}$**
Instructions$_{\text{EN}}$	INSTRUCT$_{\text{EN}}$	BAS$_{\text{EU}}$ I$_{\text{EN}}$	INSTRUCT$_{\text{EN}}$
Instructions$_{\text{EU}}$	BAS$_{\text{EN}}$ I$_{\text{EU}}$	BAS$_{\text{EU}}$ I$_{\text{EU}}$	INS$_{\text{EN}}$ I$_{\text{EU}}$
Instructions$_{\text{EN+EU}}$	BAS$_{\text{EN}}$ I$_{\text{EN+EU}}$	BAS$_{\text{EU}}$ I$_{\text{EN+EU}}$	INS$_{\text{EN}}$ I$_{\text{EN+EU}}$
Corpus$_{\text{EU}}$	BASE$_{\text{EU}}$	BASE$_{\text{EU}}$	INS$_{\text{EN}}$ C$_{\text{EU}}$
Corpus$_{\text{EU}}$ + Instructions$_{\text{EN}}$	BAS$_{\text{EN}}$ C$_{\text{EU}}$ I$_{\text{EN}}$	BAS$_{\text{EU}}$ C$_{\text{EU}}$ I$_{\text{EN}}$	INS$_{\text{EN}}$ C$_{\text{EU}}$ I$_{\text{EN}}$
Corpus$_{\text{EU}}$ + Instructions$_{\text{EU}}$	BAS$_{\text{EN}}$ C$_{\text{EU}}$ I$_{\text{EU}}$	BAS$_{\text{EU}}$ C$_{\text{EU}}$ I$_{\text{EU}}$	INS$_{\text{EN}}$ C$_{\text{EU}}$ I$_{\text{EU}}$
Corpus$_{\text{EU}}$ + Instructions$_{\text{EN+EU}}$	BAS$_{\text{EN}}$ C$_{\text{EU}}$ I$_{\text{EN+EU}}$	BAS$_{\text{EU}}$ C$_{\text{EU}}$ I$_{\text{EN+EU}}$	INS$_{\text{EN}}$ C$_{\text{EU}}$ I$_{\text{EN+EU}}$

Table 3: Summary of model variants based on different backbone LLMs and training data combinations. Each cell contains the shorthand identifier for that model variant, reflecting its configuration. Gray entries indicate redundant configurations where the backbone model has already seen the corresponding data.

Model	Hardware	# GPU	GPU hours	CO2eq
BASE$_{\text{EU}}$	64Gb H100	32	960.0h×1	105.51Kg
Any backbone + I$_{\text{EN}}$	64Gb A100	128	192.0h ×1	22.81Kg
Any backbone + I$_{\text{EU}}$	64Gb A100	128	264.5h ×3	94.26Kg
Any backbone + I$_{\text{EN+EU}}$	64Gb A100	128	456.5h ×3	162.69Kg
Any backbone + C$_{\text{EU}}$	64Gb A100	128	1,730.1h ×1	205.54Kg
Any backbone + C$_{\text{EU}}$ I$_{\text{EN}}$	64Gb A100	128	1,932.8h ×3	688.86Kg
Any backbone + C$_{\text{EU}}$ I$_{\text{EU}}$	64Gb A100	128	2,016.0h ×3	718.50Kg
Any backbone + C$_{\text{EU}}$ I$_{\text{EN+EU}}$	64Gb A100	128	2,327.5h ×3	829.53Kg
INST$_{\text{EN,70B}}$ C$_{\text{EU}}$ I$_{\text{EN}}$	64Gb A100	256	16,005.1h ×1	1,901.41Kg
Total	-	-	39,879.1h	4,729.11Kg

Table 4: Summary of training costs in GPU hours and carbon footprint (see naming conventions in Table 3)

70B model variants were trained with consistent configurations in terms of sequence length, batch size, optimizer settings, and learning rate schedules. The main differences lie in the number of GPUs and the FSDP sharding strategy, which was adjusted to better accommodate the increased memory and compute demands of the larger model. These hyperparameters were optimized using 8B model variant of BASE$_{\text{EU}}$, and the initial iteration on the instructed variants. We found this configuration to robustly perform across all the configurations.

We also explored the balance between the amount of English instructions and Basque monolingual data during joint training. We evaluated each setting on a subset of the development benchmarks, obtaining the results in Table 6, which show that the number of training epochs had a more significant impact on performance than the number of instructions. However, to better assess instruction quality, we also conducted a small-scale internal arena evaluation. We observed that models trained with 4M instructions tended to produce worse responses. Based on these findings, we selected the 1M instructions + 4 epochs configuration as the most balanced setup. We have several hypotheses which would explain the results above:

1. All examples are generated using the same method—Magpie—from the backbone model, and they tend to be quite homogeneous within their respective clusters *general*, *translation*, *code*, *math*, and *arithmetic*.

2. Prior work by Etxaniz et al. (2024b) suggests that only a limited amount of English is needed during continual pretraining to maintain cross-lingual capabilities and avoid catastrophic forgetting. That is, 1M English instructions could be enough to avoid catastrophic forgetting when doing continual pretraining with Basque data.

C Evaluation details

C.1 Static Benchmarks

Our evaluation framework comprises a total of 27 benchmarks across three languages: 14 in Basque,

Hyperparameter	8B Models	70B Models
GPUs	128	256
Sequence Length	8192	
Gradient Accumulation	1	
Micro Batch Size	2	
Total Batch Tokens	2M	4M
Epochs	4	
Optimizer	AdamW	
β_1, β_2	0.9, 0.95	
Scheduler	Cosine	
Cosine min LR ratio	0.33	
Learning rate	$1e^{-5}$	
Warm-Up ratio	0.1	
Weight Decay	0.1	
Precision	BFloat16	
FSDP Sharding Strategy	HYBRID	FULL

Table 5: Hyperparameters used to train the models

# Instructions	Epochs	Accuracy
1M	1	59.91
1M	4	61.37
4M	1	59.82
4M	4	61.97

Table 6: Average benchmark results in the preliminary hyperparameter search for number of instructions and training epochs.

9 in English, and 4 in Spanish. These benchmarks span six categories designed to test different aspects of model capabilities:

- **Reading comprehension**: Belebele (Bandarkar et al., 2024), a multilingual dataset spanning 122 languages; and EusReading (Etxaniz et al., 2024b), containing 352 complex reading comprehension exercises from official C1-level Basque examinations.
- **Common sense**: XStoryCloze (Lin et al., 2022), a multilingual version of the original StoryCloze (Mostafazadeh et al., 2017) dataset testing narrative understanding; and PIQA (Bisk et al., 2020), which assesses physical common sense through everyday tasks. PIQA's translation to Basque is available through IberoBench (Baucells et al., 2025).
- **Linguistic proficiency**: EusProficiency (Etxaniz et al., 2024b), with +5,000 questions from official Basque examinations; and BL2MP (Urbizu et al., 2024), designed to evaluate grammatical knowledge in Basque, inspired by the BLiMP benchmark methodology (Warstadt et al., 2020).
- **Miscellaneous knowledge**: BertaQA (Etxaniz et al., 2024a), which tests knowledge of local Basque culture versus global topics; EusTrivia and EusExams from the Latxa suite (Etxaniz et al., 2024b); and a subset of MMLU (Hendrycks et al., 2021), manually translated to Basque (Corral et al., 2025) and Spanish.[12]
- **Maths & Reasoning**: MGSM (Shi et al., 2023), a multilingual grade school maths benchmark; and ARC (Clark et al., 2018), for scientific reasoning. We use Basque versions of both from IberoBench.

Except MGSM, the datasets are framed as multiple-choice problems where models' answers are determined by selecting the option with the highest log probability. MGSM is implemented as a generative task where an answer is directly sampled from the evaluated model and matched against a reference answer. We specifically chose the native chain-of-thought scenario. To provide models with contextual examples, our evaluations employed a 5-shot setting.

C.2 Arena Details

Guidelines for arena participants. Fig. 11 contains the information and instructions that were given to the human annotators who participated in the community-driven arena initiative. All the data collected through this initiative was properly anonymized prior to publication. Note that the actual information was provided in Basque, while here we show a translation to English.

Arena infrastructure. On the infrastructure side, we used vLLM (Kwon et al., 2023) to serve all model pairs and the baseline. We deployed a total of 18 endpoints for 8B models and one endpoint for a 70B model, running on nine and two A100 80GB GPUs, respectively. For the frontend, we developed a lightweight Gradio (Abid et al., 2019) interface that allowed users to enter prompts, view model responses, and indicate their preferences based on content quality, language quality, or, in cases where no clear winner emerged, overall quality. To ensure

[12] hf.co/datasets/openai/MMMLU

a fair comparison across models, all models were given the same system prompt (shown in Fig. 10) and the same hyperparameters: 0.9 temperature and 0.95 top-p.

Bradley–Terry model. The Bradley–Terry model (Bradley and Terry, 1952) provides a principled probabilistic framework for aggregating pairwise preferences into a global ranking over models. Let $\mathcal{M} = \{M_1, M_2, \ldots, M_n\}$ denote the set of models under evaluation. The model assigns a latent preference strength θ_i to each model M_i. The probability that model M_i is preferred over M_j in a pairwise comparison is given by:

$$P(M_i > M_j) = \frac{e^{\theta_i}}{e^{\theta_i} + e^{\theta_j}}$$

Given a dataset $\mathcal{D}$ of observed pairwise outcomes the parameters $\{\theta_i\}$ are estimated using Maximum Likelihood Estimation (MLE). To facilitate interpretation, we apply zero-mean centering, treating the scores as deviations from the average model. The final scores for each model are then computed as:

$$Score(M_i) = 400 \cdot \theta_i + 1000$$

By using a scaling factor of 400, we ensure that the scores are interpretable in a manner consistent with the online ELO rating system. Thanks to its score stability and the assumption that model performance remains constant over time, the Bradley–Terry scoring system has become a widely adopted method for ranking LLMs—particularly since its introduction in the Chatbot Arena.[13]

C.3 Safety and Bias

We assess the extent to which our instruction-tuning strategy preserves the safety and bias alignment properties of the backbone models. Specifically, we evaluate the model variant $\text{INS}_{\text{EN}}\,\text{C}_{\text{EU}}\,\text{I}_{\text{EN}}$ in both 8B and 70B parameter sizes, comparing it with two critical counterparts: (1) $\text{BAS}_{\text{EN}}\,\text{C}_{\text{EU}}\,\text{I}_{\text{EN}}$, to analyze the impact of starting from an already instruction-tune backbone versus a base model (8B only); and (2) $\text{INSTRUCT}_{\text{EN}}$, to measure potential alignment changes introduced with our fine-tuning data mix. This evaluation aims to ensure that models maintain appropriate safety guardrails and fairness characteristics in both Basque and English.

[13] lmarena.ai

Safety. To test safety, we construct a Basque-language dataset inspired by XSTest (Röttger et al., 2024), combining clearly unsafe prompts with superficially similar but safe ones. We measure both Violation rates (VR) and False Refusal rates (FRR) where the model wrongly declines safe prompts. The dataset includes a total of 50 instances, comprising both unsafe and safe prompts across five sensitive categories (*self-harm*, *drugs*, *child-exploitation*, *terrorism*, and *explicit-content*), adapted to the Basque context and translated into English for cross-language comparison. The outputs of the models were manually annotated by three members performing red teaming. Annotation agreement for unsafe prompt outputs was high (average agreement percentage: 0.973; Fleiss' Kappa: 0.786; Krippendorff's Alpha: 0.789). For safe prompt outputs, the annotators unanimously agreed on every item.

Bias. For bias evaluation, we use BasqBBQ (Zulaika and Saralegi, 2025) for Basque and BBQ (Parrish et al., 2022) for English to analyze disparities across languages. The evaluation is conducted in the same way as described in Section 5.1, using LM Evaluation Harness framework with 4 few-shot examples. We use the accuracy metric to evaluate the bias of the model, measuring its ability to choose the correct answer even when biased traps are added to mislead it (Parrish et al., 2022).

D Detailed Results

D.1 Benchmark Results

Tables 7 and 8 present the accuracy scores for all model variants across our benchmark suite.

D.2 Arena Results

Complete human evaluation results in terms of Bradley–Terry scores and final rankings for each arena dimension (content, language, and global quality) can be consulted in Table 9. Additionally, the detailed global win-rates and battle counts for each model pair are shown in Fig. 12. Note that the reported win-rates incorporate ties in the calculation, resulting in win and loss percentages that do not necessarily sum to 100% for each model pair. Note also that base models (i.e., BASE_{EN} and BASE_{EU}) were not included in the arena evaluation, as they are not capable of following instructions. However, their performance on benchmarks can be consulted in Appendix D.1.

	ARC$_{C}$	ARC$_{E}$	Bele	BQA$_{G}$	BQA$_{L}$	BLMP	EusEx	EusPro	EusRe	EusTri	MGSM	MMLU	PIQA	XSC	Avg
BASE$_{EN}$	28.84	49.75	61.56	63.29	43.65	74.06	45.63	32.69	47.44	43.79	26.40	47.41	56.92	56.72	48.44
+ I$_{EU}$	36.35	61.95	69.44	67.39	42.98	84.06	48.39	35.42	44.89	45.42	30.80	50.37	61.44	62.14	52.93
+ I$_{EN+EU}$	38.28	63.50	72.00	68.76	42.77	84.25	48.41	36.19	45.17	46.27	37.20	**52.04**	62.09	62.84	54.27
+ C$_{EU}$ I$_{EN}$	39.42	64.81	73.00	70.99	63.49	91.28	51.92	**48.66**	**56.82**	56.33	47.20	51.85	**67.32**	**69.36**	60.89
+ C$_{EU}$ I$_{EU}$	38.14	67.00	71.78	71.24	62.65	92.39	49.88	47.20	46.02	57.78	39.20	50.00	65.63	66.71	58.97
+ C$_{EU}$ I$_{EN+EU}$	**40.10**	**68.60**	**74.89**	**72.74**	**63.79**	**92.50**	**52.78**	47.65	53.12	**59.59**	**48.40**	51.48	64.87	66.84	**61.24**
BASE$_{EU}$	38.65	67.38	**75.22**	72.45	**65.65**	**92.50**	**55.03**	**53.57**	58.24	**60.52**	36.00	51.11	65.09	68.50	**61.42**
+ I$_{EN}$	40.36	63.93	72.22	70.94	63.20	90.11	52.21	48.31	**59.09**	57.78	**45.60**	52.22	65.85	**70.55**	60.88
+ I$_{EU}$	39.25	67.09	73.56	71.78	61.59	92.11	52.38	47.78	49.43	57.14	35.60	53.70	64.92	67.31	59.55
+ I$_{EN+EU}$	38.31	66.92	74.56	**72.58**	62.23	91.22	53.04	48.96	53.98	58.95	44.80	**56.30**	64.27	67.70	60.99
+ C$_{EU}$ I$_{EN}$	**40.96**	66.12	61.22	71.49	63.83	91.72	50.67	47.19	47.16	57.49	43.60	52.59	**66.83**	70.35	59.37
+ C$_{EU}$ I$_{EU}$	37.97	67.42	71.33	71.24	62.52	91.83	51.76	46.31	52.84	57.32	37.60	52.22	65.90	67.31	59.54
+ C$_{EU}$ I$_{EN+EU}$	39.16	**68.43**	71.67	71.70	64.13	92.06	52.61	48.09	53.69	58.43	40.80	54.44	66.23	68.63	60.72
INSTRUCT$_{EN}$	29.10	50.88	73.89	67.10	44.97	69.61	46.21	34.13	49.72	45.01	45.60	50.37	57.63	61.22	51.82
+ I$_{EU}$	38.65	63.85	78.00	69.57	42.98	83.67	51.80	38.36	50.57	45.20	27.20	55.56	62.64	64.73	56.76
+ I$_{EN+EU}$	39.59	64.65	79.22	70.40	43.10	84.00	51.43	38.69	52.56	52.80	35.20	54.81	62.85	64.13	57.51
+ C$_{EU}$	37.97	65.70	77.33	73.87	**66.33**	**92.67**	55.05	52.12	58.24	61.40	48.40	51.85	66.99	70.28	62.73
+ C$_{EU}$ I$_{EN}$	**41.38**	66.79	80.00	**74.62**	65.23	91.39	56.00	**52.83**	59.66	61.05	**54.00**	**57.04**	**67.32**	**71.34**	64.19
+ C$_{EU}$ I$_{EU}$	40.44	69.15	81.44	73.54	66.07	91.83	56.01	52.06	**62.78**	**62.33**	46.40	52.96	66.01	71.01	63.72
+ C$_{EU}$ I$_{EN+EU}$	39.85	**70.16**	**83.00**	72.99	65.57	92.28	**56.23**	52.35	61.93	62.10	50.80	56.30	65.69	69.56	**64.20**
INSTRUCT$_{EN,70B}$	44.97	72.18	89.11	83.53	53.51	80.83	63.28	43.59	72.16	62.51	76.40	68.52	66.34	69.69	67.61
+ C$_{EU}$ I$_{EN}$	**55.12**	**77.57**	**91.00**	**87.42**	**77.71**	**92.11**	**71.56**	**68.00**	**78.98**	**74.17**	**80.00**	**68.89**	**70.75**	**77.83**	**76.51**
3.5 Sonnet	-	-	**94.22**	**93.52**	**80.46**	-	82.68	**81.60**	**87.39**	**84.61**	**85.20**	**79.63**	-	-	-
GPT-4o	-	-	92.89	91.01	74.83	-	79.17	74.25	84.38	80.70	79.20	76.67	-	-	-

Table 7: Accuracy scores in Basque benchmarks. Best results in each compute class are in **bold**. Best overall results are underlined.

	English										Spanish				
	ARC$_{C}$	**ARC$_{E}$**	**Bele**	**BQA$_{G}$**	**BQA$_{L}$**	**MGSM**	**MMLU**	**PIQA**	**XSC**	**Avg**	**Bele**	**MGSM**	**MMLU**	**XSC**	**Avg**
BASE$_{EN}$	54.61	84.30	**87.78**	75.59	49.11	55.20	**67.78**	80.79	81.34	69.81	81.67	50.40	57.41	74.06	65.88
+ I_{EU}	**54.95**	**83.63**	86.78	73.95	47.38	67.60	63.70	80.63	**82.46**	70.73	80.67	58.40	57.78	74.12	67.74
+ I_{EN+EU}	54.82	81.42	87.61	74.46	47.91	78.00	61.30	80.25	82.40	72.52	**82.00**	**67.60**	58.89	**75.28**	**70.94**
+ C_{EU} I_{EN}	53.84	80.72	85.89	**75.75**	60.41	**79.20**	60.00	79.71	82.40	73.03	81.33	63.60	60.00	73.53	69.62
+ C_{EU} I_{EU}	54.44	82.20	80.33	75.08	59.56	62.00	55.19	80.69	80.21	69.12	75.22	58.00	53.70	72.01	64.73
+ C_{EU} I_{EN+EU}	53.92	81.65	85.78	75.63	**60.79**	74.40	63.33	**80.85**	82.06	**73.06**	80.33	62.80	**61.11**	73.40	69.41
BASE$_{EU}$	52.39	82.83	84.33	**76.05**	**61.72**	55.20	**63.70**	80.30	80.08	68.87	79.11	46.80	58.52	72.20	64.16
+ I_{EN}	53.33	80.68	**86.44**	74.50	59.22	73.60	61.11	80.09	**81.54**	72.21	78.11	**64.40**	58.15	72.73	68.35
+ I_{EU}	54.44	82.41	84.11	74.87	58.63	64.00	62.22	80.09	81.27	70.65	77.22	56.40	54.81	74.19	65.66
+ I_{EN+EU}	53.75	80.89	85.89	74.41	59.14	73.20	**63.70**	79.22	81.14	**72.44**	**80.44**	**64.40**	56.67	**74.39**	**68.97**
+ C_{EU} I_{EN}	54.18	81.94	84.11	74.50	59.48	**74.40**	59.63	80.14	81.27	71.65	76.33	56.80	57.41	73.33	65.97
+ C_{EU} I_{EU}	53.92	**82.53**	81.78	74.50	59.09	58.00	62.96	80.09	80.21	69.09	74.89	47.60	55.56	72.60	62.66
+ C_{EU} I_{EN+EU}	**54.78**	82.24	84.56	74.58	60.62	73.20	61.48	**80.63**	81.01	71.65	76.78	60.40	**59.26**	73.20	67.41
INSTRUCT$_{EN}$	**57.76**	**85.48**	**92.67**	**77.47**	50.51	**87.20**	**66.67**	**81.28**	**83.52**	**77.02**	**87.89**	**78.80**	**62.96**	**77.50**	**76.79**
+ I_{EU}	54.18	82.07	90.89	75.38	47.88	80.00	64.81	79.60	80.54	74.09	86.44	70.00	60.37	74.85	72.92
+ I_{EN+EU}	50.85	77.99	91.56	75.59	49.11	81.60	65.19	72.25	69.49	72.10	85.89	73.20	**62.96**	74.85	74.23
+ C_{EU}	53.84	81.86	90.22	77.05	**63.16**	68.40	64.81	80.20	83.26	73.74	82.67	61.20	59.63	73.20	69.17
+ C_{EU} I_{EN}	51.96	79.42	91.00	77.38	62.01	81.20	62.59	74.59	76.31	74.35	85.67	73.20	59.26	72.47	72.65
+ C_{EU} I_{EU}	53.58	81.65	90.67	76.55	61.63	72.00	64.07	79.60	80.28	74.65	84.22	70.00	58.52	75.18	71.98
+ C_{EU} I_{EN+EU}	51.62	79.46	90.89	77.13	61.17	84.80	65.93	74.86	74.12	74.65	85.11	72.80	60.00	74.06	72.99
INSTRUCT$_{EN,70B}$	63.78	**$\underline{90.61}$**	95.44	85.49	56.98	**$\underline{94.80}$**	**$\underline{78.15}$**	**$\underline{85.04}$**	**$\underline{85.37}$**	81.74	**$\underline{94.22}$**	**$\underline{86.40}$**	77.04	**$\underline{83.92}$**	85.39
+ C_{EU} I_{EN}	**$\underline{66.89}$**	87.58	**$\underline{96.33}$**	**$\underline{88.46}$**	**$\underline{74.70}$**	**$\underline{94.80}$**	77.04	81.66	79.02	**$\underline{84.09}$**	93.33	86.00	**$\underline{84.40}$**	81.14	**$\underline{86.82}$**

Table 8: Accuracy scores in English and Spanish benchmarks. Best results in each compute class are in **bold**. Best overall results are underlined.

	Arena $_{Global}$			Arena $_{Content}$				Arena $_{Language}$			
	Rank	**Bradley–Terry**		**Rank**		**Bradley–Terry**		**Rank**		**Bradley–Terry**	
GPT-4o	1	1188	(+13/-17)	1	=	1183	(+15/-13)	1	=	1093	(+12/-10)
3.5 Sonnet	2	1153	(+13/-21)	2	=	1150	(+12/-17)	3	▼1	1082	(+11/-11)
INS $_{EN,70B}$ C $_{EU}$ I $_{EN}$	3	1141	(+15/-11)	3	=	1127	(+10/-11)	2	▲1	1083	(+13/-13)
INS $_{EN}$ C $_{EU}$ I $_{EN+EU}$	5	1050	(+13/-14)	4	▲1	1047	(+12/-12)	4	▲1	1038	(+10/-8)
INS $_{EN}$ C $_{EU}$ I $_{EU}$	5	1050	(+14/-11)	5	=	1045	(+11/-13)	6	▼1	1034	(+8/-10)
INS $_{EN}$ C $_{EU}$ I $_{EN}$	6	1038	(+13/-13)	6	=	1031	(+15/-12)	5	▲1	1036	(+11/-10)
BAS $_{EN}$ C $_{EU}$ I $_{EN+EU}$	7	1025	(+13/-11)	7	=	1026	(+13/-13)	10	▼3	1019	(+7/-12)
BAS $_{EU}$ C $_{EU}$ I $_{EN+EU}$	8	1022	(+13/-8)	8	=	1019	(+12/-11)	9	▼1	1021	(+10/-10)
BAS $_{EN}$ C $_{EU}$ I $_{EN}$	9	1017	(+12/-11)	10	▼1	1004	(+11/-14)	7	▲2	1027	(+10/-9)
BAS $_{EU}$ I $_{EN+EU}$	10	1008	(+13/-13)	9	▲1	1008	(+10/-13)	12	▼2	1008	(+9/-9)
BAS $_{EU}$ C $_{EU}$ I $_{EN}$	12	1005	(+17/-14)	13	▼1	989	(+16/-13)	8	▲4	1026	(+10/-13)
BAS $_{EU}$ I $_{EN}$	12	1005	(+12/-14)	11	▲1	1000	(+13/-13)	11	▲1	1014	(+10/-10)
BAS $_{EU}$ I $_{EU}$	13	991	(+13/-16)	12	▲1	990	(+13/-17)	15	▲2	991	(+11/-10)
BAS $_{EN}$ C $_{EU}$ I $_{EU}$	14	985	(+12/-16)	14	=	985	(+13/-15)	13	▲1	1002	(+10/-10)
BAS $_{EU}$ C $_{EU}$ I $_{EU}$	15	982	(+12/-17)	15	=	984	(+10/-14)	14	▲1	994	(+10/-11)
INS $_{EN}$ C $_{EU}$	16	942	(+9/-12)	17	▼1	935	(+8/-12)	16	=	974	(+9/-8)
INS $_{EN}$ I $_{EN+EU}$	17	936	(+14/-11)	16	▲1	943	(+10/-13)	18	▼1	944	(+11/-10)
INS $_{EN}$ I $_{EU}$	18	918	(+14/-14)	18	=	924	(+11/-12)	19	▼1	943	(+12/-11)
BAS $_{EN}$ I $_{EN+EU}$	19	915	(+16/-14)	19	=	922	(+15/-17)	17	▲2	955	(+12/-10)
BAS $_{EN}$ I $_{EU}$	20	896	(+14/-16)	20	=	909	(+12/-16)	20	=	925	(+10/-13)
INSTRUCT $_{EN}$	21	722	(+19/-17)	21	=	766	(+17/-14)	21	=	783	(+12/-12)

Table 9: Full results of human evaluation in the arena, across the three evaluation dimensions: content quality, linguistic quality and global preference. For convenience, 8B parameter-sized models are highlighted by backbone family (BASE $_{EN}$, BASE $_{EU}$, and INSTRUCT $_{EN}$). Rankings for content and language include differences (▲/▼) relative to the global ranking. Values in parenthesis indicate 90% confidence intervals for Bradley–Terry scores.

```
User:
Zein da merkatu ikerketa primarioaren eta bigarren mailakoaren arteko aldea?

Assistant:
Merkatu ikerketa primarioak jatorrizko datuak zuzenean iturritik biltzea dakar, hala nola
inkestak, foku taldeak edo esperimentuak. Ikerketa mota hori normalean garestiagoa eta denbora
gehiago behar izaten du, baina ikuspegi bakarrak eta espezifikoak eskaintzen ditu.

Merkatu ikerketa bigarren mailakoak, berriz, beste norbaitek bildutako datu existenteak
aztertzea dakar, hala nola azterketa akademikoak, industria-txostenak edo gobernu-estatistikak.
Ikerketa mota hori normalean merkeagoa eta azkarragoa da egiteko, baina ez du ziur aski lehen
mailako ikerketaren zehaztasun edo garrantzi bera eskaintzen.

Hona hemen merkatu ikerketa primarioaren eta bigarren mailakoaren arteko alde nagusiak:

**Merkatu Ikerketa Primarioa**

* Jatorrizko datuak zuzenean iturritik biltzen ditu
* Zehatzagoa da eta helburu espezifikoetara bideratuta dago
* Ikuspegi bakarrak eta espezifikoak eskaintzen ditu
* Normalean garestiagoa eta denbora gehiago behar izaten du
* Adibidez: inkestak, foku taldeak, esperimentuak, bezeroen elkarrizketak

**Merkatu Ikerketa Bigarren Mailakoa**

* Beste norbaitek bildutako datu existenteak aztertzen ditu
* Orokorragoa da eta ez da hain zehatza edo espezifikoa
* Ez du ziur aski lehen mailako ikerketaren zehaztasun edo garrantzi bera eskaintzen
* Adibidez: azterketa akademikoak, industria-txostenak, gobernu-estatistikak, lineako
berrikuspenak, sare sozialen analitika

Merkatu ikerketa primarioaren eta bigarren mailakoaren artean erabaki behar duzunean, kontuan
hartu faktore hauek:

1. **Ikerketa helburuak**: Zure xede-publikoari buruzko informazio zehatz eta xehetasunezkoa
behar baduzu, ikerketa primarioa aukera hobea izan daiteke. Orokorreko informazioa edo atzeko
planoko ikerketa behar baduzu, ikerketa bigarren mailakoarekin nahikoa izan daiteke.
2. **Aurrekontua**: Ikerketa primarioa garestia izan daiteke, bigarren mailako ikerketarekin
alderatuta.
3. **Denbora mugak**: Oro har, ikerketa bigarren mailako bat egitea errazagoa eta azkarragoa
da ikerketa primario bat egitea baino.
4. **Datuaren kalitatea**: Ikerketa primarioek jatorrizko datuak eskaintzen dituzte, bigarren
mailako ikerketek dagoeneko bildutako datuetan oinarritzen direnbitartean.

Azken finean, lehen mailako eta bigarren mailako merkatu-ikerketekin konbinazio bat emaitza
osoago bat lortzeko modurik onena izan daiteke.
```

Figure 9: The same as above (Fig. 8) but translated to Basque by 8B BASE$_{EU}$.

```
You are a helpful Artificial Intelligence assistant called [ANONYMIZED], created and developed by
[ANONYMIZED].

The user will engage in a multi-round conversation with you, asking initial questions and
following up with additional related questions. Your goal is to provide thorough, relevant and
insightful responses to help the user with their queries. Every conversation will be conducted in
standard Basque, this is, the first question from the user will be in Basque, and you should
respond in formal Basque as well. Conversations will cover a wide range of topics, including but
not limited to general knowledge, science, technology, entertainment, coding, mathematics, and
more. Today is {date}.
```

Figure 10: System prompt given to the models in the arena evaluation

Information on Data Usage

To participate in this Arena, you must provide a username and email address. This information is necessary for entry into the final raffle. All collected information will be deleted once the Arena concludes.

ATTENTION! Your username will be publicly visible throughout the Arena.

ATTENTION! Since we collect personal data and prompts/responses in this Arena, participation is restricted to individuals 14 years of age and older.

No additional personal data will be collected. However, we do collect other information, including:

- User prompts and responses.
- User preferences.

This data will be used for the following purposes:

- Evaluation of models participating in the Arena.
- Research for new models.

This data will be published openly in the future under a CC0 license. By participating in the Arena, you grant permission for this use.

Instructions for Participation

The Arena is a research initiative we've prepared at [ANONYMIZED] to help develop public chatbots for Basque. All participants will have the chance to get numbers for an amazing raffle.

Here's what you'll need to do:

- You must write and send a question or command
- Two different chatbots will respond. Your job is to analyze and compare the answers to decide which one is better. We want to measure both **content** quality and **Basque language** quality.
- In some cases, you'll be asked a third question if your content and language quality assessments are contradictory.
- After answering all questions, you'll have the option to send your assessment via the "Send evaluation" button.
- To write a new question or command, you'll need to click the "New chat" button.

To summarize, what you need to do is:

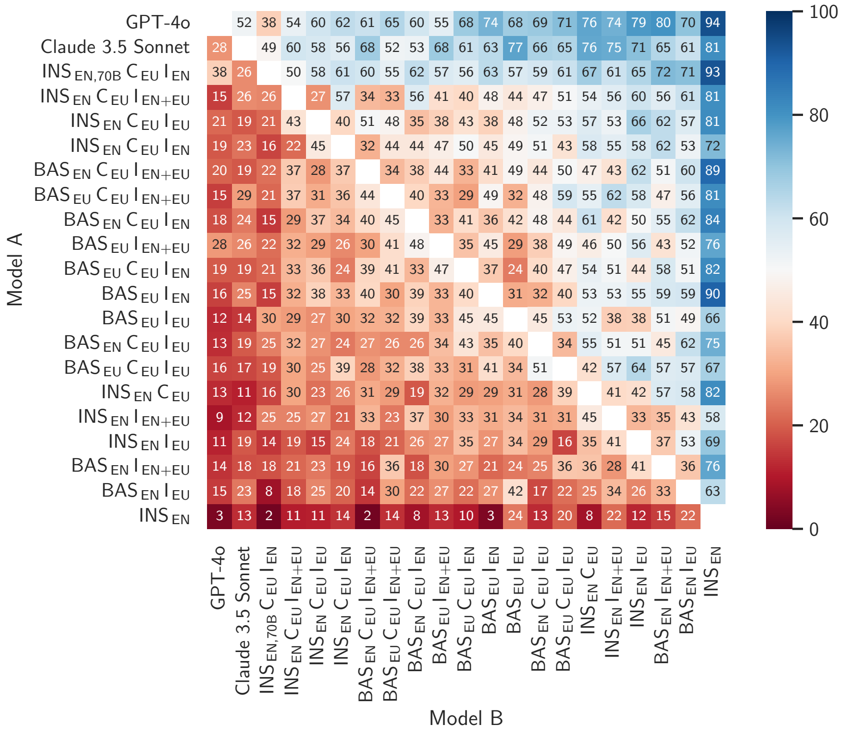
1. Write a question or command for the chatbots. For example:
    - "How do you make a potato omelet?"
    - "Summarize the following text: [...]"
2. Read both answers and compare the quality of content and quality of Basque language.
3. Decide which response you prefer in terms of content and language. For each:
    - If A is better, choose A
    - If B is better, choose B
    - If both are at the same level (good or bad), choose TIE
4. If you wish, you can continue with the conversation, ask for more explanations, or try another question. You can change your answer from step 3, taking into account the quality of the entire conversation.
5. To restart the process, click the "New chat" button.

We want your OPINION. But play fair! We will occasionally conduct an analysis of the results received and verify control answers. If they're not correct, you won't participate in the raffle.

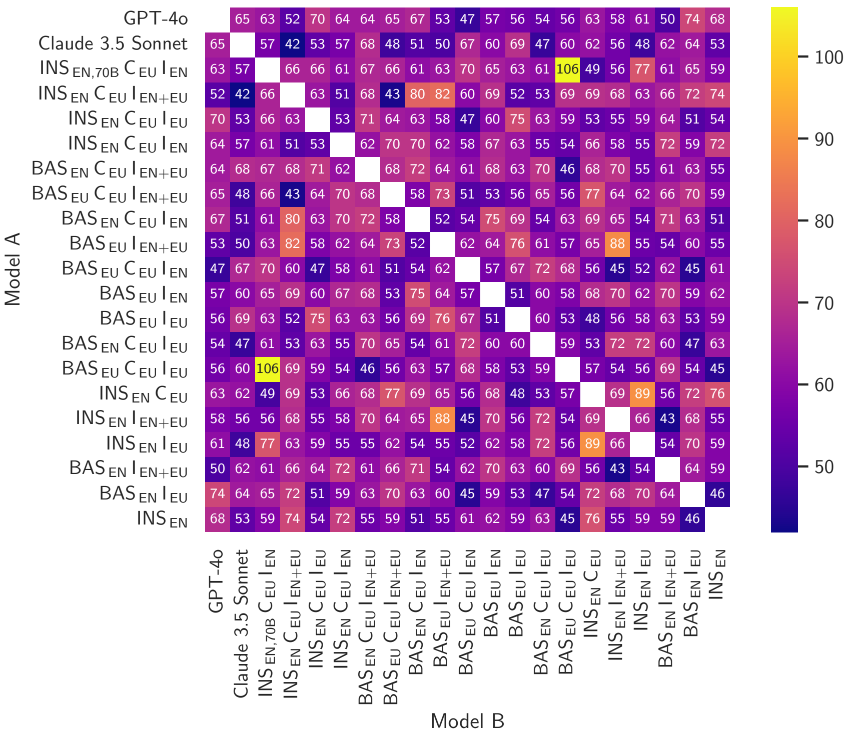
About the chatbots

In total, we've put 21 chatbots in competition. Among them are private models like GPT-4o or Claude, open models like Llama 3.1, and some we've developed ourselves. Overall, there's a variety of chatbots: good ones, very good ones, and also bad ones. In this examination, our goal is to systematically evaluate these chatbots.

Figure 11: Information panel and instructions for human participants in the arena



(a) Win-rates matrix showing the percentage of wins for Model A (row) against Model B (column)



(b) Battle counts matrix showing the number of direct comparisons performed between each model pair

Figure 12: Detailed human evaluation results from the arena study ($\text{Arena}_{\text{Global}}$)