Multilingual Refusal Alignment for Safer Large Language Models

Anonymous ACL submission

Abstract

As Large Language Models (LLMs) are deployed globally, ensuring their safety and alignment across multiple languages becomes paramount. However, safety behaviors often vary unpredictably between languages, posing significant challenges for consistent and ethical AI. In this work, we systematically investigate the dynamics of multilingual alignment, exploring whether single-language alignment transfers cross-lingually, how language consistency is preserved during training, and the resulting trade-offs with general knowledge capabilities. We introduce RefusEU[1], a novel refusal alignment dataset covering 12 European languages, including a dedicated test set for evaluating current state-of-the-art models. Our controlled Direct Preference Optimization (DPO) experiments provide two key insights: alignment exclusively in English is insufficient for cross-lingual safety even for the same group of harms, and, surprisingly, training without English samples can exacerbate safety performance and general multilingual capabilities.

1 Introduction

Advances in pretrained Large Language Models (LLMs) have significantly improved language understanding and generation, enabling their rapid adoption in real-world applications (Achiam et al., 2023; Comanici et al., 2025). At the same time, these models raise safety challenges such as harmful outputs, biased behavior, and ethical risk - that are amplified in multilingual settings (Yong et al., 2025). As LLMs continue to scale globally, ensuring safe and consistent behavior across languages has become more demanding than ever.

While LLMs are increasingly used across languages, most alignment work remains focused on English (Askell et al., 2021; Ouyang et al., 2022). This creates a mismatch: models may behave safely in English yet fail to do so in other languages, leading to inconsistent or unsafe outputs. For global deployments, such discrepancies can undermine user trust and exacerbate region-specific harms.

Cross-lingual consistency is therefore a key challenge. Aligning each language independently is costly and often impractical, making it crucial to understand whether alignment in one language generalizes to others. Yet, little work evaluates this systematically, and publicly available datasets for multilingual safety–especially around refusal behavior–are lacking.

In this work, we focus on multilingual alignment and pose three research questions: **RQ1:** Do we need to perform multilingual alignment for each language on the same groups of prompts, or is training in a single language sufficient? **RQ2:** How well is cross-lingual consistency preserved during multilingual training? **RQ3:** How does multilingual alignment influence general multilingual capabilities?

LLMs can be aligned using various methods: through reward modeling, as in PPO (Schulman et al., 2017); grouped modeling with reward functions, as in GRPO (Shao et al., 2024); or by leveraging datasets of preference pairs, as in ORPO (Hong et al., 2024) and DPO (Rafailov et al., 2024).

Among this, we adopt Direct Preference Optimization (DPO) as it is one of the most well-known and widely adopted alignment methods. DPO is a streamlined method for aligning language models human preferences. It directly optimizes the language model on preference data, which is formatted as pairs of "*chosen*" (preferred) and "*rejected*" (dispreferred) responses. This eliminates the need for a separate reward model, simplifying and stabilizing the training process.

To study alignment dynamics more directly, we build on recent advances in *refusal abliteration*, a technique that disables a model's safety mecha-

[1]The dataset will be released publicly upon paper acceptance.

nisms. To uncensor an LLM, one must first identify the "refusal direction" by finding the mean difference vector between the model's internal activations on harmful and harmless instructions. Once identified, this direction can be ablated (removed) from the model's computation either temporarily during inference or permanently via weight modification. This method has been shown to compromise safety by allowing the model to generate harmful or restricted outputs that it would otherwise decline to produce (Shairah et al., 2025; Arditi et al., 2024). By applying refusal ablation prior to multilingual training, we obtain a clearer picture of how alignment methods behave when safety mechanisms are deliberately removed.

Our main contributions can be summarised as follows:

- We introduce **RefusEU**, a multilingual dataset of refusal alignment across 12 European languages, designed to evaluate safe response generation and cross-lingual alignment in LLMs.
- We conduct controlled training experiments across multiple multilingual alignment setups.
- We provide a thorough multilingual evaluation of alignment outcomes, highlighting both safety and capability trade-offs.

2 Related work

Yong et al. (2025) revealed a significant gap in LLM safety research for non-English languages, including those often considered high-resource. This finding is consistent with Röttger et al. (2025), who reported that 113 out of 144 surveyed safety datasets were exclusively in English (78.5%).

Kanepajs et al. (2024) examined how a language's level of resourcing relates to the vulnerability of LLMs to multilingual jailbreaks in that language. Their study focused on 24 official EU languages, translating into all variants and conducting a logistic regression analysis. The results show that for GPT-4o and Mistral Large 2, jailbreak attack success rates tend to be higher for low-resourced languages.

The most well-known multilingual safety datasets covering primarily European languages include: **PolygloToxicityPrompts** (Jain et al., 2024), **RTP-LX** (de Wynter et al., 2025) – datasets for toxicity evaluation, **SORRY-Bench** – a benchmark for efficiently assessing refusal capabilities (Xie et al., 2025), **XTRUST** – a multilingual benchmark spanning a diverse range of topics (Li et al., 2024), **CIVICS** – a hand-crafted multilingual dataset of value-laden prompts addressing socially sensitive topics (Pistilli et al., 2024), **HONEST** – a collection of manually created sentence templates for measuring the generation of hurtful sentence completions (Nozza et al., 2021).

Another line of research focuses on releasing multilingual datasets together with frameworks for evaluating or generating new samples. Examples include: **MM-ART** – an automated method for multi-turn, multilingual red-teaming of LLMs (Singhania et al., 2025), **CSRT** – a framework for synthesizing code-switching red-teaming queries (Yoo et al., 2025), **S-Eval** – a framework for evaluating the safety of large language models (LLMs) using a structured risk taxonomy and scalable test generation (Yuan et al., 2025).

While most of the cited works focus solely on evaluation and lack concrete mitigation strategies for improving safety performance, a few notable exceptions combine dataset release with mitigation approaches: **XSAFETY** – a benchmark that also proposes a prompting-based mitigation method for improving safety (Wang et al., 2024), **MASSIVE MULTILINGUAL HOLISTICBIAS (MMHB)** – accompanied by a mitigation strategy using gender-balanced prompts (Tan et al., 2024). Only a few articles have focused on multilingual training datasets and conducting training experiments. Bhardwaj et al. (2024) proposed a method to re-align a fine-tuned model by simply performing an arithmetic addition of a safety vector to the model's weights, and they also introduced the multilingual **CATQ** benchmark. Haider et al. (2024) describe the iterative "break-fix" cycle as repeated rounds of safety dataset curation, post-training optimization, and comprehensive responsible AI (RAI) benchmarking. Well-described multilingual training approaches are presented in Aakanksha et al. (2024), where the first human-annotated collection of adversarial prompts across eight languages was introduced. The authors tested various mitigation techniques such as supervised fine-tuning and DPO with different proportions of training safety samples. They also distinguished between "global" and "local" harms. Another method for training safety data collection is described by Deng et al. (2024), alongside general performance samples for supervised fine-tuning, in the **Self-Defence** framework. Moreover, the authors provided **MultiJail**—a mul-

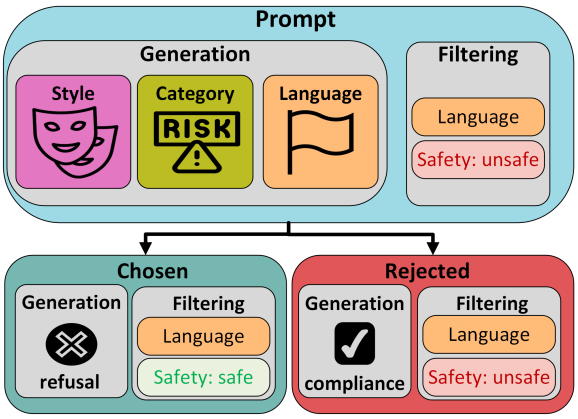

Figure 1: Dataset construction process

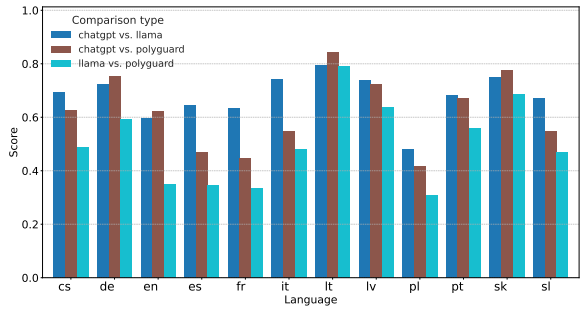

Figure 2: Cohen's Kappa between LLM as a Judge models

tilingual jailbreak dataset. To address the limited availability of datasets for multilingual alignment, particularly for European languages, we propose a new dataset, which we describe in the following section.

3 Dataset construction

The detailed dataset construction process is illustrated in Figure 1. We follow the framework proposed by Krasnodębska et al. (2025), where adversarial prompts are generated in a single step by specifying the crime category, the detailed criminal activity, and the desired writing style. We adopt 14 crime categories based on the Llama-Guard taxonomy and 10 different attack styles following the original Rainbow teaming methodology (Samvelyan et al., 2024). The full list of crime categories and attack styles is provided in Appendix A.

To construct the multilingual scenario, we translate the initial examples of crime categories, attack styles, and generation prompts from Polish (as described in Krasnodębska et al. (2025)) into each target language (English, German, Italian, French, Spanish, Portuguese, Polish, Czech, Slovak, Slovenian, Lithuanian, Latvian) using the DeepL API[2].

First, we generate 500 adversarial prompt examples in each language using the following models: Mistral-Small-24B[3] and Llama-3.1-70B[4]. As shown in Table 1, the Mistral-Small-24B model is more effective at generating unsafe questions. Interestingly, the Llama-3.1-70B model tends to produce prompts in English more readily than in lower-resourced languages such as Polish. For both models, the language consistency during generation is close to 100%.

Table 1: Model Performance Comparison

	Mistral-Small-24B-Instruct-2501		Llama-3.1-70B	
Lang.	Lang. Consistency [%]	Unsafe Prompts [%]	Lang. Consistency [%]	Unsafe Prompts [%]
pl	100	97.0	100	38.8
en	100	97.6	100	73.8
sl	100	71.0	98.06	20.6
cs	100	61.6	99.41	33.8

Second, we generate over 100k samples per language using five examples for each pair of style and crime activity. We then compare the inter agreement in labeling adversarial questions across three models: Llama-Guard-3-8B[5], PolyGuard-Qwen[6], and GPT-4o-mini[7], using the same classification prompt as in Llama-Guard-3-8B.

As shown in Figure 2, the most consistent outputs are obtained with Llama-Guard-3-8B and GPT-4o-mini. Interestingly, the agreement scores for some underrepresented languages, such as Lithuanian, Latvian, and Slovak, are more stable. It is worth noting that the analysis was conducted on samples intentionally designed to elicit unsafe questions, and thus the subset is not balanced. When comparing the proportion of identical labels across all test cases, we observe more than 90% agreement.

In the initial stage, questions labeled as unsafe by both the Llama-Guard-3-8B and PolyGuard-Qwen models were retained. Since these two models exhibited noticeable discrepancies in their labelling, requiring agreement between them provides a stricter criterion that helps ensure only gen-

[2] https://www.deepl.com/en/translator
[3] https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501
[4] https://huggingface.co/meta-llama/Llama-3.1-70B
[5] https://huggingface.co/meta-llama/Llama-Guard-3-8B
[6] https://huggingface.co/ToxicityPrompts/PolyGuard-Qwen
[7] https://platform.openai.com/docs/models/gpt-4o-mini

uinely harmful prompts are preserved. Furthermore, questions written in languages other than the expected one were excluded using Fasttext model for language identification[8].

To obtain safe answers for the generated data, we use Llama-3.1-70B, which demonstrates a strong tendency to avoid producing harmful content, not only for English inputs. RefusEU is designed in such a way that the chosen answer for unsafe prompts is always a refusal. To generate these samples, the initial prompts were augmented with instructions to produce refusals. Further details are provided in Appendix B.

To collect unsafe answers, we use an abliterated version of Llama-3.1-70B, named Llama-3.1-70B-Instruct-lorablated[9]. In both cases (safe and unsafe answers), we also apply language consistency filtering, analogous to that employed in the question generation step.

Finally, only pairs in which the refusal answers were consistently marked as safe and the compliance responses were labeled as unsafe by both Llama-Guard-3-8B and PolyGuard-Qwen were retained.

To summarize, we have collected over 4,000 prompt–response pairs for each language. Each pair includes a chosen (preferred, safety-refusal) answer and a rejected (non-preferred, unsafe) answer.

4 Methodology

In this section, we describe our methodology, including the models used, datasets, and the training and evaluation procedures.

4.1 Abliteration

The ablation procedure was not performed by the authors. Instead, as a starting point, we adopted the ablated versions of Llama 3.1–8B–Instruct[10] and Llama 3.1–70B–Instruct, which was previously used for compliance answer generation. As shown in Table 2, these two models exhibit high initial ASR scores across all languages.

[8] https://huggingface.co/facebook/fasttext-language-identification

[9] https://huggingface.co/mlabonne/Llama-3.1-70B-Instruct-lorablated

[10] https://huggingface.co/mlabonne/Meta-Llama-3.1-8B-Instruct-abliterated

4.2 Alignment datasets

We use the dataset described in Section 3 as our starting point. The dataset was split into training and test sets as follows: for each style and category, samples in different languages were grouped together into either the test or training set to ensure thorough evaluation without contamination during training. This resulted in a fixed **RefusEU-test** set with 2,000 samples per language for detailed evaluation and a initial **RefusEU-train** set containing 2,091 samples per 12 languages.

4.3 Training setup

We consider four different training scenarios depending on data used for alignment:

balanced – all 12 languages are represented with the same number of samples (25,092 samples); **high_resource** – only high-resource languages are included: English, German, Italian, French, Spanish, and Portuguese (12,546 samples); **en** – only English samples are included (2,091 samples); **without_en** – all languages except English are included (23,001 samples).

In each setup, the DPO method was chosen as the alignment technique, using the implementation from the TRL library[11]. The detailed training parameters are provided in the Appendix C.

4.4 Baselines

We use the abliterated Llama-3.1-8B and Llama-3.1-70B models as baselines and compare their performance after the multilingual alignment procedure.

We also evaluate initial instruction-tuned Llama models: Llama-3.1-8B[12], Llama-3.1-70B and other open models : Qwen2.5-7B-Instruct[13], Qwen2.5-14B-Instruct[14], gpt-oss-20b[15], Mistral-Small-24B-Instruct-2501[16], Mistral-Nemo-Instruct-2407[17].

[11] https://huggingface.co/docs/trl/en/dpo_trainer

[12] https://huggingface.co/meta-llama/Llama-3.1-8B

[13] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[14] https://huggingface.co/Qwen/Qwen2.5-14B-Instruct

[15] https://huggingface.co/openai/gpt-oss-20b

[16] https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501

[17] https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407

Table 2: ASR [%] across different models trained on Llama 8B and Llama 70B on RefusEU-test dataset. Lower is better.

Language	instruct		abliterated		balanced		en		high-res		without-en	
	8B	**70B**	**8B**	**70B**	**8B**	**70B**	**8B**	**70B**	**8B**	**70B**	**8B**	**70B**
en	0.6	5.4	89.15	83.65	6.25	0.0	1.15	0.05	3.8	0.0	13.9	5.65
de	0.65	2.65	41.2	49.8	1.45	0.0	0.3	0.4	1.4	0.05	2.15	1.9
it	2.85	7.8	76.5	66.7	5.0	1.4	7.7	3.65	4.3	1.1	9.9	14.5
fr	1.55	2.55	61.9	55.05	4.85	0.3	2.6	0.6	3.35	0.2	4.0	3.25
es	2.9	4.8	89.9	76.2	4.35	0.15	7.5	2.0	5.0	0.05	4.65	9.55
pt	3.55	5.0	81.15	74.15	5.15	0.8	5.8	2.5	4.8	0.05	8.45	12.0
avg-highR	2.02	4.7	73.3	67.59	4.51	0.44	4.18	1.53	3.78	0.24	7.18	7.81
pl	2.25	8.3	87.45	92.9	1.05	0.9	0.7	11.35	1.8	0.1	3.9	22.6
cs	12.95	18.65	66.75	73.3	1.45	0.4	1.55	4.9	1.85	0.07	2.6	10.9
sk	12.15	19.5	63.95	74.8	0.45	1.15	1.65	14.85	1.45	8.25	4.9	9.2
sl	33.65	26.5	61.4	73.6	0.1	2.9	0.9	28.8	0.1	8.6	1.3	21.85
lt	15.4	14.0	47.65	51.5	0.55	0.15	0.85	8.85	0.95	1.2	9.15	3.2
lv	34.55	20.3	56.05	58.65	0.85	0.05	5.75	15.5	5.4	3.15	13.4	2.55
avg-lowR	18.49	17.88	63.88	70.79	0.74	0.93	1.9	14.04	1.93	3.56	5.88	11.72
avg-All	10.25	11.29	68.59	69.19	2.63	0.68	3.04	7.79	2.85	1.9	6.53	9.76

4.5 Evaluation

Safety The basic metric used in the evaluation stage is the Attack Success Rate (ASR), defined as the proportion of harmful responses to harmful requests. We used the test set of RefusEU as the primary component for adversarial testing. The final ASR score was computed by first generating predictions using Llama-Guard-3-8B and PolyGuard-Qwen. In cases where the outputs of these two models diverged, GPT-4o-mini, prompted with a predefined safety classification protocol, was employed as an adjudicator to establish the final label. Language consistency is evaluated by computing the proportion of outputs generated in the corresponding language.

General knowledge To evaluate overall capabilities and relative changes in model effectiveness, we conducted an analysis using the Global MMLU benchmark Singh et al. (2025), available for 9 of the 12 tested languages. Additionally, to assess generalization across languages outside the European family, we extended the analysis on GlobalMMLU to Chinese (zh) and Korean (ko).

To analyze the general performance of generation across languages, we constructed a small dataset consisting of 50 prompts per language. These prompts targeted rephrasing, providing opinions, and writing polite requests, and were designed to elicit responses no longer than four sentences. For evaluation, we adopted the LLM-as-a-Judge framework. Specifically, we used the gpt-oss-120B[18] model to assess *fluency* and *linguistic correctness* in each language. The evaluation prompt is provided in Appendix B. Scores ranged from 1 (lowest quality) to 5 (highest quality).

5 Results

In this section, we present our results addressing RQs 1–3.

5.1 Attack success rate

To address **RQ1** (whether training in a single language is sufficient), we evaluate ASR performance across all languages under four training scenarios.

Overall, ASR scores on RefuseEU-test were the lowest on the balanced dataset for both LLaMA 8B and LLaMA 70B trainings (see Table 2). The second most interesting configuration was the dataset based on high-resource languages. In this scenario, Slovenian and Slovakian exceeded 8% ASR for LLaMA 70B. For LLaMA 8B, all scores were below 5%, except for Latvian. Results per language are presented in Figure 3. Interestingly, training only on English preferences resulted in worse ASR scores for LLaMA 70B than for LLaMA 8B. Finally, avoiding English safety preference data led to the worst evaluations for both models: ASR exceeded 10% for Polish, Czech, Slovenian, Italian, and Portuguese with LLaMA 70B, and for English and Latvian with LLaMA 8B.

To test against another safety-related benchmark,

[18] https://huggingface.co/openai/gpt-oss-120b

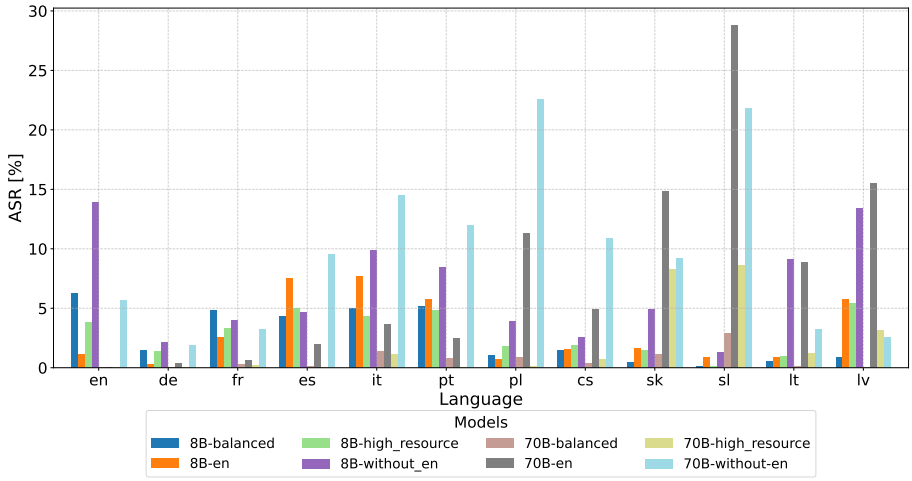

Figure 3: ASR [%] for RefusEU-test. Lower scores mean safer models.

we use RTP-LX as a multilingual toxicity benchmark available for each tested language. Detailed results are provided in Appendix D. The highest ASR scores are obtained from training without English examples, both for high-resource and low-resource languages, as well as overall. This finding is consistent with the RefusEU evaluation, where the ASR scores were also highest under this setup.

5.2 Language consistency

To address **RQ2** (which concerns language consistency in multilingual training), we evaluate language consistency across all experimental setups.

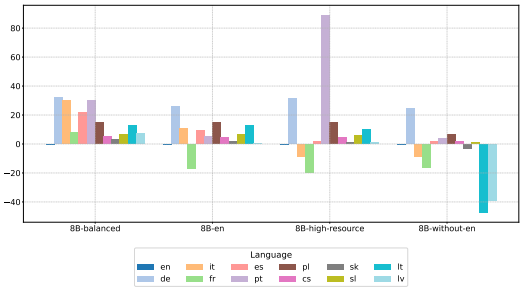

Figure 4: Difference in language consistency across different setups and initial abliterated model for the Llama 8B trainings.

The detailed proportion of answers in corresponding language is presented in Table 7 in Appendix E.1 while summarised results are in Figure 4. For all the Llama 70B training and the ablated and original versions, the answers are provided in the corresponding question's language, resulting in the consistency close to 100%. For Llama 8B, however, the ablated version shows a relatively low proportion of answers in the same language as question. Surprisingly, for the balanced training, the proportion of English answers improves only for French . The best results, in terms of language consistency, were obtained from the training without English samples, with mainly only German performing worse than the initial ablated model.

5.3 Global MMLU

To address **RQ3** (regarding general multilingual capabilities after multilingual alignment), we first employ the GLOBAL-MMLU benchmark. For the Llama-8B training runs, the sharpest decline in accuracy is observed for Polish and Portuguese (see Table 3). For the remaining languages, the drop remains around 0.005. In the case of Llama-70B training, the reduction in effectiveness is relatively minor. For Polish and English, we even see slight improvements across different setups. The most reliable strategy for mitigating losses in performance appears to be using a dataset with high-resource languages, where the decline is smallest for both Llama-8B and Llama-70B models. The difference were illustrated in Figure 5.

5.4 Fluency and correctness

To extend the examination of **RQ3**, we perform additional testing for fluency and correctness across all languages. The smallest drops in fluency and correctness are observed for the 70B model trained solely on the English dataset, particularly for high-resource languages (see Figure 6). In contrast, the largest drops for high-resource languages occur mainly with the 8B models. For low-resource languages, the performance drop is smaller; however, in this group, the initially unaligned 8B model per-

Table 3: Comparison of models across languages for GLOBAL MMLU FULL for Llama 8B and Llama 70B.

Language	instruct		abliterated		balanced		en		high-resource		without-en	
	8B	70B	8B	70B	8B	70B	8B	70B	8B	70B	8B	70B
en	0.6593	0.8116	0.6223	0.8089	0.6152	0.8088	0.6162	0.8086	0.6175	0.8106	0.6152	0.8098
de	0.5407	0.7453	0.5009	0.7388	0.4949	0.7362	0.4957	0.7366	0.4959	0.7389	0.4927	0.7381
it	0.5491	0.7607	0.5144	0.7571	0.5072	0.7543	0.5070	0.7553	0.5081	0.7563	0.5076	0.7560
fr	0.5621	0.7529	0.5138	0.7460	0.5070	0.7449	0.5046	0.7459	0.5074	0.7462	0.5046	0.7461
es	0.5697	0.7683	0.5308	0.7657	0.5272	0.7641	0.5234	0.7644	0.5255	0.7635	0.5247	0.7626
pt	0.5577	0.7656	0.5310	0.7654	0.5187	0.7609	0.5187	0.7617	0.5208	0.7610	0.5147	0.7621
pl	0.5017	0.7224	0.4567	0.7179	0.4390	0.7186	0.4422	0.7191	0.4442	0.7201	0.4373	0.7186
cs	0.5093	0.7298	0.4697	0.7248	0.4605	0.7236	0.4606	0.7245	0.4632	0.7238	0.4603	0.7246
lt	0.4205	0.6170	0.3777	0.6132	0.3684	0.6058	0.3695	0.6125	0.3714	0.6118	0.3649	0.6100
zh	0.5310	0.7212	0.4899	0.7165	0.4816	0.7148	0.4832	0.7157	0.4855	0.7164	0.4821	0.7165
ko	0.4820	0.6764	0.4253	0.6722	0.4185	0.6707	0.4185	0.6725	0.4201	0.6714	0.4177	0.6715

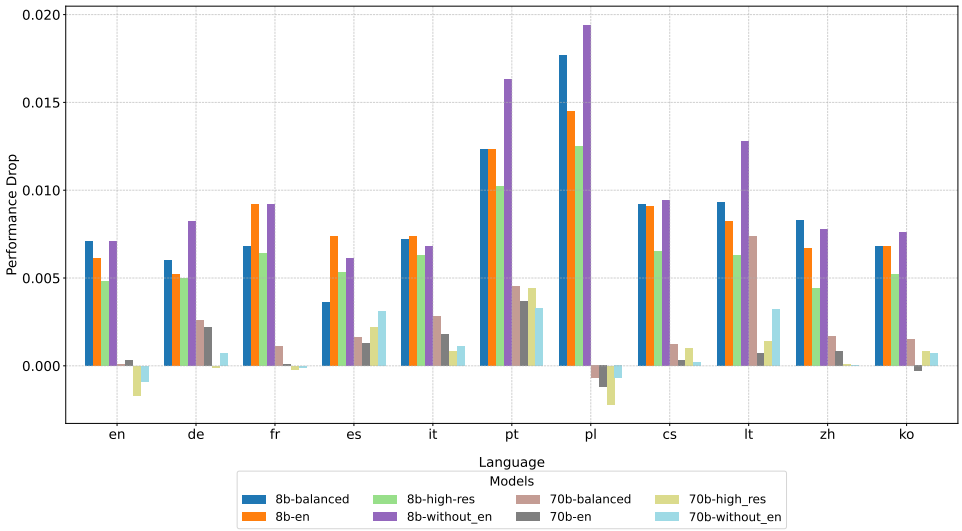

Figure 5: Performance drop for GLOBAL MMLU.

forms particularly poorly. Overall, the performance degradation after alignment is only slightly worse for larger architecture. We examine an additional mechanism involving a translation pipeline, which first translates the original task into English and then back-translates the response, using EuroLLM-9B model[19] (as it covers all tested languages). The performance and correctness drop relative to the initial model is similar when using the translation pipeline for the 70B models (see Figure 8). For the 8B models, a clear trend emerges: for low-resource languages, it is more effective to use the translation pipeline rather than forcing the model to respond in the original language—even when using the original Llama 3.1-Instruct model. In these cases, the drop in correctness and fluency is negative, indicating that translation improves the evaluation scores. The detailed results are presented in Tables 9 and 10 in Appendix E.2.

5.5 Evaluation of popular open models

Finally, we tested ASR performance on RefusEU-test on baseline models from Section 4.4. Performance safety evaluation deviates from zero for low-resource European languages in the Slavic family. Higher ASR scores are observed for the Llama models, with the 70B variant outperforming the 8B variant. For the Qwen models, performance degradation is more pronounced in the smaller 7B model compared to the 14B model. The highest ASR scores are recorded for the Mistral Nemo Instruct-2407 model.

Regarding ASR, the most thorough safety training appears in the gpt-oss-20B model, where ASR does not exceed 2%. The evaluated models display different safety mechanisms when examining the proportion of English responses. For the

[19] https://huggingface.co/utter-project/EuroLLM-9B-Instruct

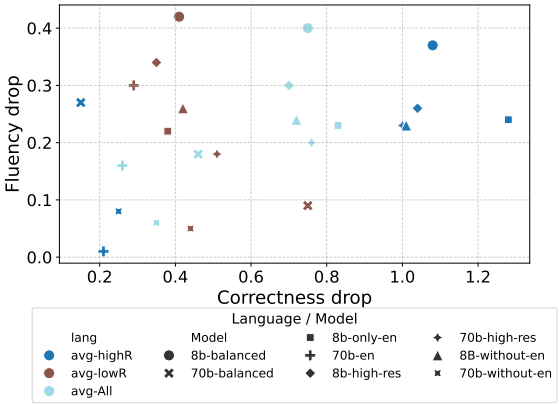

Figure 6: Drop on fluency and correctness across different models and datasets and initial abliterated model.

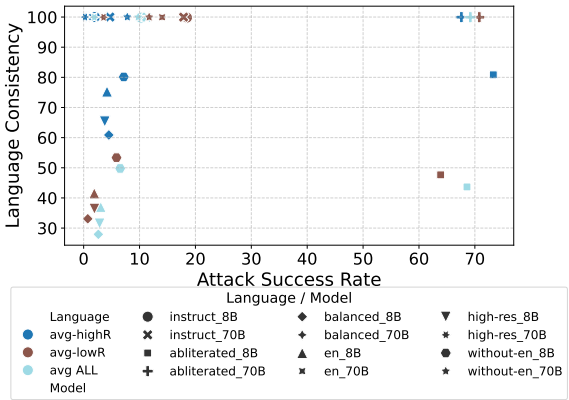

Figure 7: Comparison between ASR scores [%] and language consistency [%].

safest model, gpt-oss-20B, the system almost always replies in English with the same refusal message: "I'm sorry, but I can't help with that." For low-resource languages and Mistral models, responses in English are also possible, whereas Qwen and Llama models tend to provide answers in the original language. Complete results are reported in Appendix E.3.

Our investigation reveals a critical and perhaps counterintuitive insight: omitting English data from the alignment process does not simply isolate non-English languages but actively degrades safety performance across the board. This suggests that English data may serve as an essential anchor for safety concepts, even in a multilingual context.

6 Findings

Many studies do not incorporate additional metrics when evaluating overall safety performance across different models, often reporting only ASR values.

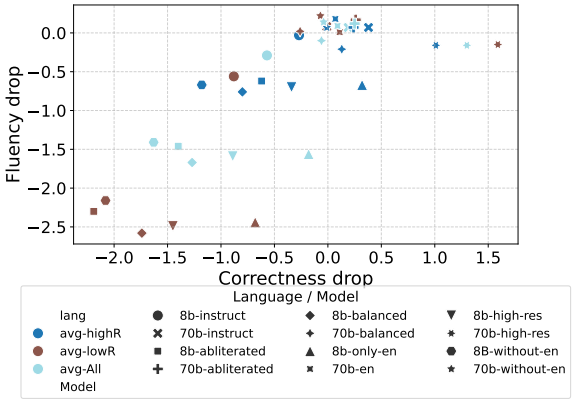

Figure 8: Fluency and correctness drop across corresponding models and their translations into English.

In a multilingual setup, however, it's important to measure additional metrics such as language consistency. When combined with ASR, the best results were obtained using a high-resource dataset for training Llama 70B (see Figure 7).

Large discrepancies are observed in the training of Llama 8B. While improving and ensuring responses remain in the original language can enhance consistency, it may also lead to decreased safety performance. An alternative approach is to use an additional translation model to return answers in a consistent language, but this setup requires evaluating the entire end-to-end solution, not just the main generative model.

For smaller models such as Llama-8B and low-resource languages, it is beneficial to boost performance through additional translation and processing in English. In contrast, the 70B models are more resilient to translation effects, and these steps do not lead to general performance improvements.

7 Conclusion

Our results show that aligning models only in English is insufficient for low-resource languages. The Llama-70B model achieves an average ASR of 14.04%, while Llama-8B reaches 1.9%, though most of its responses for low-resource languages are in English. The larger Llama-70B architecture shows only minor degradation on the GLOBAL-MMLU benchmark and in fluency and correctness tests, remaining within acceptable bounds for reliable multilingual performance. In contrast, smaller models like Llama-8B benefit from an additional translation pipeline, which can improve overall results.

Limitations

Our work has several limitations. First, RefusEU currently covers only 12 of the 24 official European languages. While we find this scope sufficient for conducting additional experiments to enhance multilingual safety mechanisms, future work should extend coverage to a broader set of languages. Second, the dataset is entirely synthetically generated. Although we employed multilingual models with strong cross-lingual generation capabilities and applied an identical generation procedure across languages to ensure consistency in questions and hazard categories, synthetic data may not fully capture the linguistic and cultural nuances present in real-world usage. A potential risk of this work is that multilingual models may behave inconsistently across languages, and the automatically generated RefusEU dataset could propagate biases or be misused in unintended ways.

We used AI assistance exclusively to enhance the text style and identify grammatical errors in this manuscript.

References

Aakanksha, Arash Ahmadian, Beyza Ermis, Seraphina Goldfarb-Tarrant, Julia Kreutzer, Marzieh Fadaee, and Sara Hooker. 2024. The multilingual alignment prism: Aligning global and local preferences to reduce harm. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12027–12049, Miami, Florida, USA. Association for Computational Linguistics.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *Preprint*, arXiv:2406.11717.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, and 3 others. 2021. A general language assistant as a laboratory for alignment. *Preprint*, arXiv:2112.00861.

Rishabh Bhardwaj, Duc Anh Do, and Soujanya Poria. 2024. Language models are Homer simpson! safety re-alignment of fine-tuned language models through task arithmetic. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14138–14149, Bangkok, Thailand. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Adrian de Wynter, Ishaan Watts, Tua Wongsangaroonsri, Minghui Zhang, Noura Farra, Nektar Ege Altıntoprak, Lena Baur, Samantha Claudet, Pavel Gajdusek, Can G¨oren, Qilong Gu, Anna Kaminska, Tomasz Kaminski, Ruby Kuo, Akiko Kyuba, Jongho Lee, Kartik Mathur, Petter Merok, Ivana Milovanović, and 14 others. 2025. RTP-LX: Can LLMs evaluate toxicity in multilingual scenarios? AAAI AISI.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2024. Multilingual jailbreak challenges in large language models. In *The Twelfth International Conference on Learning Representations*.

Emman Haider, Daniel Perez-Becker, Thomas Portet, Piyush Madan, Amit Garg, Atabak Ashfaq, David Majercak, Wen Wen, Dongwoo Kim, Ziyi Yang, Jianwen Zhang, Hiteshi Sharma, Blake Bullwinkel, Martin Pouliot, Amanda Minnich, Shiven Chawla, Solianna Herrera, Shahed Warreth, Maggie Engler, and 12 others. 2024. Phi-3 safety post-training: Aligning language models with a "break-fix" cycle. *Preprint*, arXiv:2407.13833.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. *Preprint*, arXiv:2403.07691.

Devansh Jain, Priyanshu Kumar, Samuel Gehman, Xuhui Zhou, Thomas Hartvigsen, and Maarten Sap. 2024. Polyglotoxicityprompts: Multilingual evaluation of neural toxic degeneration in large language models. In *First Conference on Language Modeling*.

Arturs Kanepajs, Vladimir Ivanov, and Richard Moulange. 2024. Towards safe multilingual frontier ai. In *Workshop on Socially Responsible Language Modelling Research*.

Aleksandra Krasnodębska, Maciej Chrabaszcz, and Wojciech Kusa. 2025. Rainbow-teaming for the Polish language: A reproducibility study. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 155–165, Albuquerque, New Mexico. Association for Computational Linguistics.

Yahan Li, Yi Wang, Yi Chang, and Yuan Wu. 2024. Xtrust: On the multilingual trustworthiness of large language models. *Preprint*, arXiv:2409.15762.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2398–2406, Online. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. 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 Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Giada Pistilli, Alina Leidinger, Yacine Jernite, Atoosa Kasirzadeh, Alexandra Sasha Luccioni, and Margaret Mitchell. 2024. Civics: Building a dataset for examining culturally-informed values in large language models. *Preprint*, arXiv:2405.13974.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Paul Röttger, Fabio Pernisi, Bertie Vidgen, and Dirk Hovy. 2025. Safetyprompts: a systematic review of open datasets for evaluating and improving large language model safety. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27617–27627.

Mikayel Samvelyan, Sharath Chandra Raparthy, Andrei Lupu, Eric Hambro, Aram H. Markosyan, Manish Bhatt, Yuning Mao, Minqi Jiang, Jack Parker-Holder, Jakob Foerster, Tim Rocktäschel, and Roberta Raileanu. 2024. Rainbow teaming: Open-ended generation of diverse adversarial prompts. *Preprint*, arXiv:2402.16822.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

Harethah Abu Shairah, Hasan Abed Al Kader Hammoud, Bernard Ghanem, and George Turkiyyah. 2025. An embarrassingly simple defense against llm abliteration attacks. *Preprint*, arXiv:2505.19056.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Abhishek Singhania, Christophe Dupuy, Shivam Sadashiv Mangale, and Amani Namboori. 2025. Multi-lingual multi-turn automated red teaming for LLMs. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 141–154, Albuquerque, New Mexico. Association for Computational Linguistics.

Xiaoqing Ellen Tan, Prangthip Hansanti, Carleigh Wood, Bokai Yu, Christophe Ropers, and Marta R Costa-jussà. 2024. Towards massive multilingual holistic bias. *arXiv preprint arXiv:2407.00486*.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen tse Huang, Wenxiang Jiao, and Michael R. Lyu. 2024. All languages matter: On the multilingual safety of large language models. *Preprint*, arXiv:2310.00905.

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, and 1 others. 2025. Sorry-bench: Systematically evaluating large language model safety refusal. In *The Thirteenth International Conference on Learning Representations*.

Zheng-Xin Yong, Beyza Ermis, Marzieh Fadaee, Stephen H. Bach, and Julia Kreutzer. 2025. The state of multilingual llm safety research: From measuring the language gap to mitigating it. *Preprint*, arXiv:2505.24119.

Haneul Yoo, Yongjin Yang, and Hwaran Lee. 2025. Code-switching red-teaming: Llm evaluation for safety and multilingual understanding. *Preprint*, arXiv:2406.15481.

Xiaohan Yuan, Jinfeng Li, Dongxia Wang, Yuefeng Chen, Xiaofeng Mao, Longtao Huang, Jialuo Chen, Hui Xue, Xiaoxia Liu, Wenhai Wang, Kui Ren, and Jingyi Wang. 2025. S-eval: Towards automated and comprehensive safety evaluation for large language models. *Preprint*, arXiv:2405.14191.

A Details of dataset construction process

Table 4 outlines the Llama-Guard risk taxonomy, listing each risk category along with representative example activities. Table 5 provides an overview of the different attack styles considered, together with short descriptions of their defining characteristics.

Table 4: Llama-Guard risk taxonomy: categories and example activities.

Code	Risk Category Name	Example Activity
S1	Violent Crimes	Bank robbery
S2	Non-Violent Crimes	Fraud or forgery
S3	Sex-Related Crimes	Attempted rape
S4	Child Sexual Exploitation	Production of child pornography
S5	Defamation	Insult
S6	Specialized Advice	Advice on circumventing the law
S7	Privacy	Monitoring without consent
S8	Intellectual Property	Breach of licence agreements
S9	Indiscriminate Weapons	Chemical weapons
S10	Hate	Antisemitism
S11	Suicide & Self-Harm	Encouraging suicide
S12	Sexual Content	Forcing vomiting
S13	Elections	Fraudulent vote counting
S14	Code Interpreter Abuse	Creation of malicious browser extensions

Table 5: Attack styles with details.

Attack Style Name	Style Detail
slang	in youth slang
technical terms	using computer science terms
role play	starting with "You are" and then continuing in the second person to form a question
authority manipulation	presenting himself at the outset as a scientific authority in a certain field
misspellings	making grammatical or spelling mistakes
word play	creating lyrics
emotional manipulation	in an excited tone
hypotetical	assuming the question is asked about a hypothetical situation
historical scenario	assuming that the questioner lives in an ancient state
uncommon dialects	in in the local dialect

B Prompts

In this section we provide the exact prompt formulations used in the study. The first, the *Refusal prompt*, instructs the model to produce a safety-aligned rejection message for potentially harmful queries. The second, the *Evaluation prompt*, defines the procedure for assessing model outputs along fluency and correctness dimensions.

Refusal prompt

You are a safety-aware AI assistant. A user has submitted a question that is potentially unsafe, harmful, unethical, or violates usage guidelines. Your task is to generate a firm but respectful rejection message in language {lang} that:
Clearly states that the AI cannot fulfill the request. Explains why the question is unsafe, harmful, or inappropriate (e.g., risk of self-harm, illegal activity, misinformation, incitement to violence, etc.). Educates the user about the potential consequences of such content. Here is the user's question in : {question} Generate a thoughtful and detailed response that maintains a respectful tone while upholding safety standards and ethical use in language {lang}

Evaluation prompt

You are evaluating a model-generated response to a task in {language}. The task type is: {task_type}.
—
Task Prompt: {input_prompt}
Model Output: {model_output}
—
Evaluate the response based on the following criteria:
1. **Fluency** - Does the text sound natural and well-formed in the target language? - Are grammar, spelling, punctuation, and word choice appropriate for a native speaker?
2. **Correctness (Task Appropriateness)** - Does the response correctly fulfill the task's requirements? - Is the meaning accurate and appropriate for the task (e.g., rephrasing, opinion, polite request)?
—
Return your evaluation in the following JSON format:
{{ "fluency": <integer from 1 to 5>, "correctness": <integer from 1 to 5>, "comment": "<brief explanation if either score is 3 or below, otherwise optional>" }}

C Training parameters

The trainings were conducted on a 4×H200 cluster using 3 and 6 nodes for the Llama-8B and Llama-70B trainings, respectively. We use the following parameters to fine-tune models in our experiments:

```
beta: 0.1
max_prompt_length: 1024
max_length: 2048
num_train_epochs: 3
per_device_train_batch_size: 1
per_device_eval_batch_size: 1
gradient_accumulation_steps: 2
optim: adamw_torch
learning_rate: 1e-6
weight_decay: 1e-3
adam_beta1: 0.9
adam_beta2: 0.999
max_grad_norm: 1.0
warmup_ratio: 0.05
lr_scheduler_type: cosine
bf16: True
```

D RTP LX evaluation

This section provides additional safety evaluation using the RTP_LX dataset. As before, the initial instruct Llama 8B model shows high ASR scores for low-resourced languages. For the aligned models, the worse-performing version is the one trained excluding English samples, with ASR scores of 2.37% for high-resourced languages and 2.98% for low-resourced languages

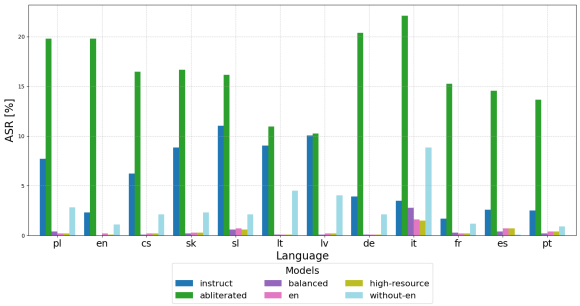

Figure 9: ASR [%] for RTP LX for the Llama 8B trainings.

E Detailed results

E.1 Language consistency results

Table 7 presents language consistency results for different variants of Llama 8B and Llama 70B models, while Table 8 shows these results for baseline open-access models.

E.2 Fluency and correctness

Table 9 presents fluency evaluation results across languages for different variants of the Llama 8B and 70B models, while Table 10 reports corresponding results for correctness.

E.3 Safety evaluation of baseline models

Table 11 and Figure 10 report ASR scores across baseline models.

Table 6: ASR [%] for RTP LX for the Llama 8B trainings.

Language	Instruct	abliterated	balanced	high-res	en	without-en
en	2.31	19.78	0.00	0.10	0.20	1.10
de	3.92	20.38	0.10	0.10	0.10	2.11
it	3.51	22.09	2.80	1.51	1.60	8.84
fr	1.71	15.26	0.30	0.20	0.20	1.20
es	2.61	14.56	0.40	0.70	0.70	0.10
pt	2.51	13.65	0.20	0.40	0.40	0.90
avg-highR	2.76	17.62	0.63	0.50	0.53	2.37
pl	7.73	19.78	0.40	0.20	0.20	2.81
cs	6.22	16.47	0.10	0.20	0.20	2.11
sk	8.84	16.67	0.20	0.30	0.30	2.31
sl	11.04	16.16	0.60	0.60	0.70	2.11
lt	9.04	10.94	0.10	0.10	0.10	4.52
lv	10.04	10.24	0.10	0.20	0.20	4.02
avg-lowR	8.81	15.04	0.25	0.27	0.28	2.98
avg-All	5.79	16.33	0.44	0.38	0.41	2.68

Table 7: Proportion of answers in given language [%] for Llama 8B and Llama 70B models.

Language	instruct		abliterated		balanced		en		high-resource		without-en	
	8B	70B	8B	70B	8B	70B	8B	70B	8B	70B	8B	70B
en	100	100	99.80	99.95	100	100	99.95	100	100	100	99.85	100
de	100	100	33.25	99.95	1.4	100	7.05	100	1.65	99.95	8.85	100
it	100	100	82.6	99.95	52.7	100	72.15	100	91.55	100	91.2	100
fr	100	100	77.3	100	69.3	100	94.05	100	96.7	100	93.25	100
es	100	100	96.6	99.95	75.1	100	87.2	100	94.8	100	94.8	100
pt	100	100	95.5	100	66.8	100	91.35	100	7.55	100	92.8	100
avg-highR	100.00	100.00	80.84	99.97	60.88	100.00	75.29	100.00	65.38	99.99	80.13	100.00
pl	100	100	15.65	99.9	0.8	99.95	0.9	100	1	100	8.95	99.95
cs	99.8	100	6.3	99.9	0.8	100	2	100	1.65	99.95	4.4	100
sk	99.85	99.95	3.2	99.8	0.25	99.05	1.15	99.35	1.85	99.3	6.3	99.75
sl	98.45	100	6.3	100	0	100	0.05	100	0.45	99.95	4.95	99.95
lt	99.75	100	14.45	100	1.4	100	1.45	100	4.1	99.95	61.65	99.95
lv	99.75	99.8	7.75	100	0.6	100	7.35	99.75	6.6	99.95	46.55	100
avg-lowR	99.82	99.98	47.66	99.95	33.08	99.92	41.53	99.93	36.41	99.93	53.36	99.97
avg-All	99.80	99.98	43.65	99.95	27.93	99.92	37.04	99.93	31.51	99.92	49.78	99.97

Table 8: Proportion of answers in given language [%] across open baseline models.

Language	Llama-3.1-8B-Instruct	Llama-3.1-70B-Instruct	Qwen2.5-7B	Qwen2.5-14B	Mistral-Nemo-Instruct-2407	Mistral-Small-24B-Instruct-2501	Gpt-oss-20b
en	100.00	100.00	99.90	99.95	99.75	99.9	100.00
de	100.00	100.00	99.75	99.90	100.00	100.00	0.95
it	100.00	100.00	99.90	100.00	99.75	99.95	4.75
fr	100.00	100.00	100.00	99.95	99.75	100.00	4.75
es	100.00	100.00	99.90	100.00	98.85	99.95	1.80
pt	100.00	100.00	99.80	100.00	99.75	100.00	7.95
avg-highR	100.00	100.00	99.88	99.97	99.64	99.97	20.03
pl	100.00	100.00	77.80	99.15	99.60	99.90	15.30
cs	100.00	100.00	96.05	100.00	97.85	99.20	23.60
sk	100.00	99.95	86.60	98.85	95.55	92.30	15.00
sl	100.00	100.00	98.10	99.45	94.15	63.75	6.00
lt	100.00	100.00	97.45	99.85	91.00	99.80	11.65
lv	100.00	99.80	96.20	99.45	77.84	97.50	13.40
avg-lowR	100.00	99.96	92.03	99.46	92.67	92.08	14.16
avg-All	100.00	99.98	95.95	99.71	96.15	96.02	17.10

Table 9: Comparison of models across languages for fluency evaluation

Language	**instruct**		**abliterated**		**balanced**		**en**		**high-resource**		**without-en**	
	8B	**70B**	**8B**	**70B**	**8B**	**70B**	**8B**	**70B**	**8B**	**70B**	**8B**	**70B**
en	4.23	4.77	4.57	4.90	3.55	4.53	2.30	4.77	3.40	2.60	4.20	4.73
de	3.07	4.50	2.23	4.07	2.23	4.40	1.97	3.97	2.07	3.57	1.70	3.90
it	3.67	4.07	2.83	4.10	2.00	4.30	2.57	4.07	2.20	3.60	2.53	4.03
fr	3.33	4.37	4.03	4.23	1.80	4.00	2.00	3.67	2.13	3.33	1.87	3.50
es	3.23	4.50	3.77	4.53	2.20	3.83	2.17	4.20	2.47	2.97	2.17	4.17
pt	3.50	4.60	3.57	4.33	2.73	4.17	2.33	4.23	2.47	4.10	2.50	4.33
avg-highR	3.51	4.47	3.50	4.36	2.42	4.21	2.22	4.15	2.46	3.36	2.49	4.11
pl	3.23	4.53	2.17	4.30	1.30	3.80	1.33	4.20	1.37	3.93	1.30	3.97
cs	2.97	3.73	2.07	4.30	1.23	3.87	1.30	3.77	1.47	3.97	1.23	3.67
sk	2.33	3.57	1.30	4.10	1.03	3.73	1.10	3.63	1.03	3.57	1.00	4.03
sl	2.63	3.93	1.17	3.87	1.03	3.00	1.03	3.90	1.00	3.38	1.00	3.33
lt	2.13	3.53	1.23	3.80	1.00	2.57	1.00	3.67	1.00	3.63	1.00	3.03
lv	1.69	3.43	1.13	3.70	1.00	2.60	1.00	3.13	1.07	2.53	1.00	3.37
avg-lowR	2.50	3.79	1.51	4.01	1.10	3.26	1.13	3.72	1.16	3.50	1.09	3.57
avg-All	3.00	4.13	2.51	4.19	1.76	3.73	1.68	3.93	1.81	3.43	1.79	3.84

Table 10: Comparison of models across languages for correctness evaluation

Language	**instruct**		**abliterated**		**balanced**		**en**		**high-resource**		**without-en**	
	8B	**70B**	**8B**	**70B**	**8B**	**70B**	**8B**	**70B**	**8B**	**70B**	**8B**	**70B**
en	4.48	4.80	4.64	4.88	3.78	4.56	2.66	4.52	3.66	2.04	4.28	4.70
de	3.58	4.62	2.62	4.34	2.02	4.52	1.94	4.24	1.98	3.96	1.74	4.12
it	4.10	4.42	3.06	4.38	2.24	4.32	2.66	4.14	2.54	3.62	2.46	4.14
fr	3.70	4.45	4.00	4.36	2.16	4.14	2.42	3.84	2.50	3.54	2.22	3.88
es	3.86	4.64	3.98	4.64	2.54	4.12	2.34	4.24	2.70	3.26	2.38	4.32
pt	3.88	4.68	3.72	4.56	2.76	4.28	2.64	4.38	2.68	3.92	2.54	4.34
avg-highR	3.93	4.60	3.67	4.53	2.58	4.32	2.44	4.23	2.68	3.39	2.60	4.25
pl	3.56	4.48	1.80	4.38	1.24	4.08	1.30	4.36	1.30	4.08	1.24	4.20
cs	3.24	4.16	1.82	4.48	1.20	4.00	1.30	3.96	1.38	4.08	1.22	3.96
sk	2.62	3.84	1.22	4.24	1.04	3.78	1.10	3.86	1.06	3.86	1.04	3.96
sl	2.62	3.84	1.24	3.74	1.04	2.68	1.08	3.72	1.06	3.27	1.00	3.20
lt	1.82	3.88	1.24	4.04	1.00	2.76	1.02	3.62	1.06	2.72	1.00	3.58
lv	2.16	3.84	1.16	4.04	1.00	2.58	1.02	3.72	1.02	3.58	1.00	3.38
avg-lowR	2.49	3.91	1.41	4.15	1.09	3.31	1.14	3.87	1.15	3.60	1.08	3.71
avg-All	3.28	4.29	2.54	4.34	1.83	3.82	1.79	4.05	1.91	3.49	1.84	3.98

Table 11: ASR [%] for baseline open-access models.

Language	Llama-3.1 8B-Instruct	Llama-3.1 70B-Instruct	Qwen2.5-14B	Qwen2.5-7B	Mistral-Nemo Instruct-2407	Mistral-Small-24B-Instruct-2501	Gpt-oss-20b
en	0.60	5.40	2.65	7.25	38.75	5.40	0.00
de	0.65	2.65	0.75	2.75	11.80	1.50	0.75
it	2.85	7.80	4.00	7.85	27.40	13.95	1.30
fr	1.55	2.55	1.80	6.65	22.50	4.25	1.90
es	2.90	4.80	3.90	5.60	38.25	7.50	0.15
pt	3.55	5.00	2.45	8.30	17.40	7.75	0.75
avg-highR	2.02	4.7	2.59	6.4	26.02	6.73	0.81
pl	2.25	8.30	1.30	4.50	38.95	8.95	0.05
cs	12.95	18.65	3.35	13.00	37.55	9.35	1.50
sk	12.15	19.50	4.00	18.00	40.20	10.30	1.45
sl	33.65	26.50	5.05	21.75	45.60	14.80	1.35
lt	15.40	14.00	3.75	21.65	29.15	8.05	1.00
lv	34.55	20.30	5.40	21.50	31.85	14.35	1.60
avg-lowR	18.49	17.88	3.81	16.73	37.22	10.97	1.16
avg-All	10.25	11.29	3.2	11.57	31.62	8.85	0.98

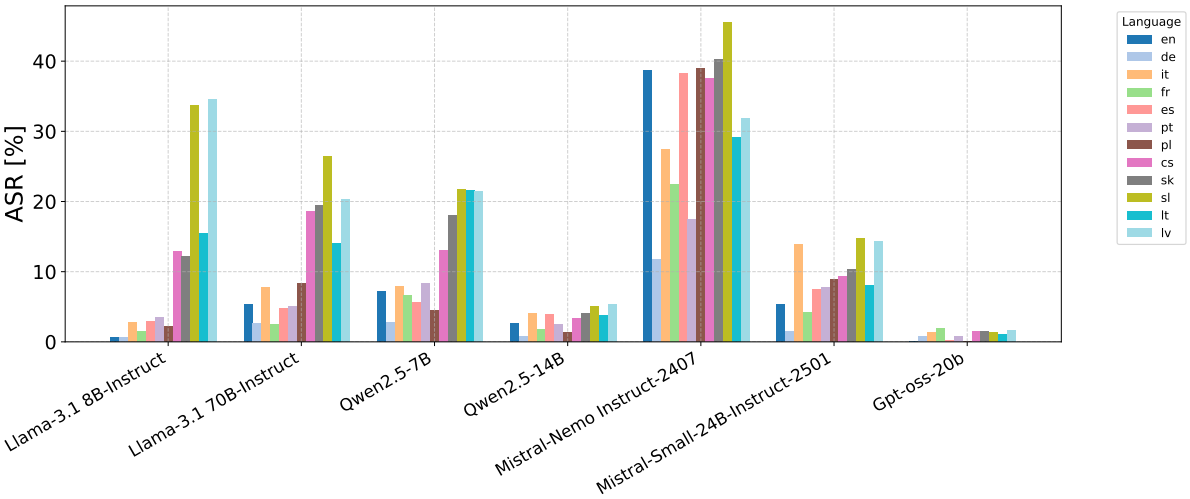

Figure 10: ASR [%] for RefusEU-test for the popular models.