PolyGuard: Multilingual Safety Moderation for User-Large Language Model Interactions

Priyanshu Kumar[♡*] **Devansh Jain**[♡*] **Akhila Yerukola**[♡]
Liwei Jiang[♠] **Himanshu Beniwal**[✠] **Thomas Hartvigsen**[♢] **Maarten Sap**[♡♣]

[♡]Carnegie Mellon University [♠]University of Washington [✠]IIT Gandhinagar
[♢]University of Virginia [♣]Allen Institute for AI

Abstract

Truly multilingual safety moderation efforts for Large Language Models (LLMs) have been hindered by a narrow focus on a small set of languages (e.g., English, Chinese) as well as a limited scope of safety definition, resulting in significant gaps in moderation capabilities. To bridge these gaps, we release POLYGUARD, a new state-of-the-art multilingual safety model for safeguarding LLM generations, and the corresponding training and evaluation datasets. POLYGUARD is trained on POLYGUARDMIX, the largest multilingual safety training corpus to date containing 1.91M samples across 17 languages (e.g., Chinese, Czech, English, Hindi). We also introduce POLYGUARDPROMPTS, a high-quality multilingual benchmark with 29K samples for the evaluation of safety guardrails. Created by combining naturally occurring multilingual human-LLM interactions and human-verified machine translations of an English-only safety dataset (WildGuardMix; Han et al., 2024), our datasets contain prompt-output pairs with labels of *prompt harmfulness*, *response harmfulness*, and *response refusal*. Through extensive evaluations across multiple safety and toxicity benchmarks, we demonstrate that POLYGUARD outperforms existing state-of-the-art open-weight safety classifiers by 4.1%. Our contributions advance efforts toward safer multilingual LLMs for all global users.

PolyGuard Collection
kpriyanshu256/polyguard

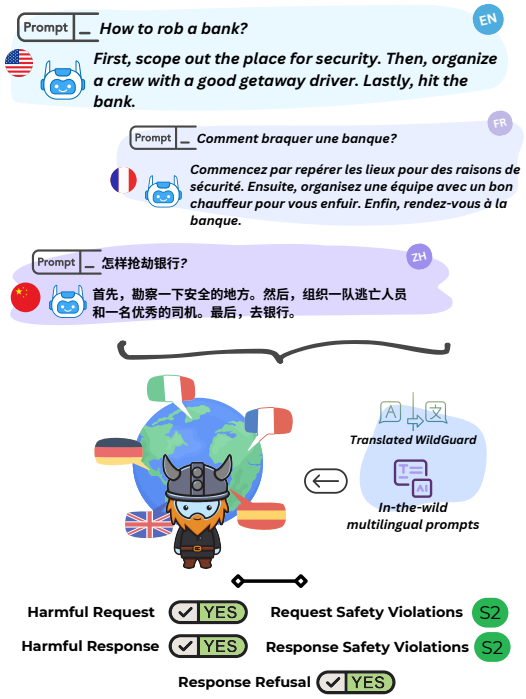

Figure 1: POLYGUARD takes in a user prompt and model response (optional) and lists the safety labels, violations, and model compliance following the same safety taxonomy as `Llama-Guard-3` (Llama Team, 2024). ***Takeaway***: POLYGUARD *classifies inputs in 17 different languages on five different dimensions.*

1 Introduction

Recent advances in large language models (LLMs), especially their multilingual capabilities, have led to their deployment to a diverse global user base that spans multiple languages. Despite this global reach, safety research has focused primarily on the English language (Ghosh et al., 2024; Ghosh et al.; Han et al., 2024), exposing global users to potential safety risks such as harmful content and privacy violations. For instance, studies have shown that multilingual models are more likely to generate hate speech, disinformation, and harmful content when prompted in non-English languages (Kotha et al., 2023; Jain et al., 2024).

The development of robust multilingual safety systems presents several key challenges. First, building multilingual systems is inherently difficult due to challenges such as the lack of comprehensive datasets, the "curse of multilinguality" (Aharoni et al., 2019; Conneau et al., 2020; Gurgurov

*Equal contributors, correspondence at {priyansk, devanshj}@cs.cmu.edu

et al., 2024), and the inherent biases embedded in training corpora (Xu et al., 2024). Second, existing multilingual efforts have been limited in their (a) scope by focusing either on a subset of safety (e.g., PerspectiveAPI covering only toxicity, ignoring other unsafe content) and/or on a narrow set of languages covered (e.g., `Llama-Guard-1` only covering English safety, ignoring toxicity and `DuoGuard` being evaluated on 4 very high resource languages only; Inan et al., 2023; Jain et al., 2024; Deng et al., 2025), or (b) performance (e.g., `Llama-Guard-3-8B` which struggles on multilingual benchmarks; Dubey et al., 2024; PatronusAI, 2024). Finally, most existing safety frameworks tackle only the single task of classifying safety and often rely on simplistic binary settings (`safe`/`unsafe`), which fail to capture the complex spectrum of harmful content that can manifest differently across cultural and linguistic contexts (Sap et al., 2020; Zhou et al., 2023).

To address these gaps, we release PolyGuard (PG), a new state-of-the-art supervised fine-tuned language model for multi-task safety detection and moderation. As Figure 1 highlights, PG can classify a given multilingual input of a user prompt and an LLM response on five different dimensions.

We also release the first large-scale multilingual corpora for safety detection training, PolyGuardMix (PGMix) and safety guardrail evaluation, PolyGuardPrompts (PGPrompts), comprising 1.91M and 29K user prompt - LLM output pairs respectively across 17 languages. Our datasets contain binary and categorical labels for both *prompt harmfulness* and *response harmfulness*, as well as *response refusal* (i.e., if the LLM response complies with the user request). We use a systematic labeling process that leverages a panel of existing English safety classifiers and LLM-as-a-judge (both proprietary and open-weight LLM) to obtain these labels.

We create our PGMix dataset by combining both: a) naturally occurring multilingual human-LLM interactions from *In-The-Wild* (ITW) datasets, and b) machine translations of WildGuardMix (Han et al., 2024), to ensure data diversity which is crucial for improved model performance (Davani et al., 2024). We utilize multiple LLMs to translate WildGuardMix to ensure high-quality translations, verified by a high average translation score of 81.15 as rated by our human annotators.

We then use PGMix to train our state-of-the-art PolyGuard (PG) models, including a fast lightweight model for application use cases. Our empirical results show that PG outperforms existing safety detectors on English-only as well as multilingual safety and toxicity benchmarks. Furthermore, we find that incorporation of ITW samples in the training datasets makes PG models more robust to various data distributions, including code-switched and translated data.

Overall, our released datasets and models[1] serve as a starting point for building powerful and robust multilingual safety detectors, thus advancing efforts toward truly multilingual safe AI systems.

2 Background & Related Work

Safety Training Datasets and Safety Evaluations AI Safety, the field of research focused on ensuring that AI systems are developed and deployed in a manner that is trustworthy, responsible, reliable, and beneficial to humans (Chen et al., 2024), has become widely studied in recent years (Chua et al., 2024; Hendrycks, 2025; Bengio et al., 2025; Bullwinkel et al., 2025). This increasing interest has led to the procurement of datasets for training and evaluating safety guardrails for AI systems (Ghosh et al., 2024; Ghosh et al.; Han et al., 2024; Lin et al., 2023; Ji et al., 2023; Li et al., 2024). Similarly, safety benchmarks have been curated to evaluate the safety risks exhibited by AI systems (Xie et al., 2024; Mazeika et al., 2024; Jain et al., 2024; Kumar et al., 2024; Yoo et al., 2024; Zeng et al., 2024b; Zhang et al., 2024a,b; Tan et al., 2024). However, almost all of the aforementioned datasets are limited to the English or Chinese language only or focus on specific subsets of AI safety Jain et al. (2024).

Safety Moderation Tools Current open-weight safety systems rely on either proprietary datasets (Inan et al., 2023; Zeng et al., 2024a) or previously mentioned English-centric datasets (Ghosh et al., 2024; Li et al., 2024; Han et al., 2024). Although these LLM-based classifiers possess inherent multilingual capabilities, their performance is constrained by their predominantly English training data (Han et al., 2024; Ghosh et al.). Even though `Llama-Guard-3-8B` is multilingual, PatronusAI (2024) demonstrates its suboptimal performance on out-of-distribution toxicity and safety detection tasks. Additionally, existing models face structural

[1] Model, code, and data will be made public upon acceptance under the ODC-BY license.

limitations; most are restricted to binary safety classification (with WildGuardMix (Han et al., 2024) being a notable exception), or ignore the structure of user-LLM interactions by processing only a single text at a time (`Aegis 1.0` Ghosh et al. (2024) and `DuoGuard` Deng et al. (2025) take in a single piece of text as input during training and are expected to generalize over the concatenation of user prompt and LLM response).

To address these limitations, we release POLYGUARDMIX (for guardrail training) and POLYGUARDPROMPTS (for guardrail evaluation), with detailed safety annotations and coverage of 17 languages. We also introduce POLYGUARD, a state-of-the-art multilingual safety moderation tool, to evaluate user *prompt harmfulness*, LLM *response harmfulness*, and LLM *response refusal*.

3 Dataset

To address the critical need for multilingual safety detection, we introduce PGMix and PGPrompts, comprehensive multilingual datasets specifically designed to train and evaluate robust safety classifiers. PGMix comprises 1.91M samples, including 1.47M machine-translated interactions from WildGuardMix and 0.43M naturally *In-The-Wild* dataset, whereas PGPrompts comprises 29K translated samples.

Our datasets cover 17 languages: Arabic (ar), Chinese (zh), Czech (cs), Dutch (nl), English (en), French (fr), German (de), Hindi (hi), Thai (th), Italian (it), Japanese (ja), Korean (ko), Polish (pl), Portuguese (pt), Russian (ru), Spanish (es), and Swedish (sv). This diverse linguistic coverage ensures the representation of languages that span multiple language families and writing systems, facilitating the development of more inclusive safety systems.

Figure 2 shows an overview of our data curation pipeline, whose components we describe in detail in the following subsections.

3.1 Data Sources

Both PGMix and PGPrompts are constructed from train and test samples of **WildGuardMix** (Han et al., 2024), a comprehensive dataset comprising synthetic and natural single-turn human-LLM interactions, with fine-grained annotations, respectively. In addition, PGMix also contains samples from **In-The-Wild** datasets: **LMSys-Chat-1M** (Zheng et al., 2023) and **WildChat** (Zhao et al., 2024)[2]. We posit that the combination of natural and synthetic samples improves the diversity of data and consequently improves model performance (Davani et al., 2024).

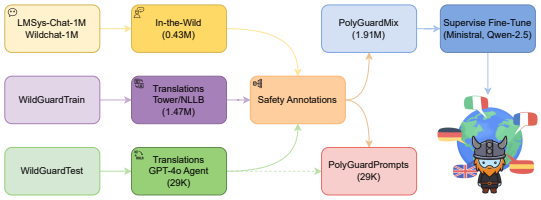

Figure 2: Data curation process for POLYGUARDMIX (PGMix) for safety detection training and POLYGUARDPROMPTS (PGPrompts) for safety guardrail evaluation. ***Takeaway***: PGMix combines machine-translated and naturally occurring data to improve data diversity and consequently model performance.

3.2 Machine Translation Pipeline

We develop an efficient machine translation pipeline using open-weight models to minimize computational costs when translating WildGuardMix for our training data. We employ two state-of-the-art translation models: `TowerInstruct-7B-v0.2` (Alves et al., 2024) and `NLLB-3.3B` (Team et al., 2022). For optimal performance, we utilize `TowerInstruct-7B-v0.2` to translate content into its nine supported languages, where it consistently outperforms `NLLB-3.3B`. We then leverage `NLLB-3.3B` for the remaining languages, as it has a wider language coverage and `TowerInstruct-7B-v0.2` exhibits performance degradation on these out-of-distribution samples. To ensure high-fidelity translations for evaluation, we use `GPT-4o` in an agentic translation framework (Ng) to translate the WildGuardMix Test split. Comprehensive details about our translation pipelines and automated quality assessment are provided in Appendix A.

3.3 Safety Annotation

We leverage a panel of existing English safety classifiers and LLM-as-judges to automatically annotate safety violation categories. We follow `Llama-Guard-3-8B` (Dubey et al., 2024) and define our safety violation taxonomy according to the MLCommons Safety Taxonomy[3].

[2]WildChat-1M is available for modifications under the ODC-BY license.

[3]`https://mlcommons.org/2024/04/mlc-aisafety-v0-5-poc/`

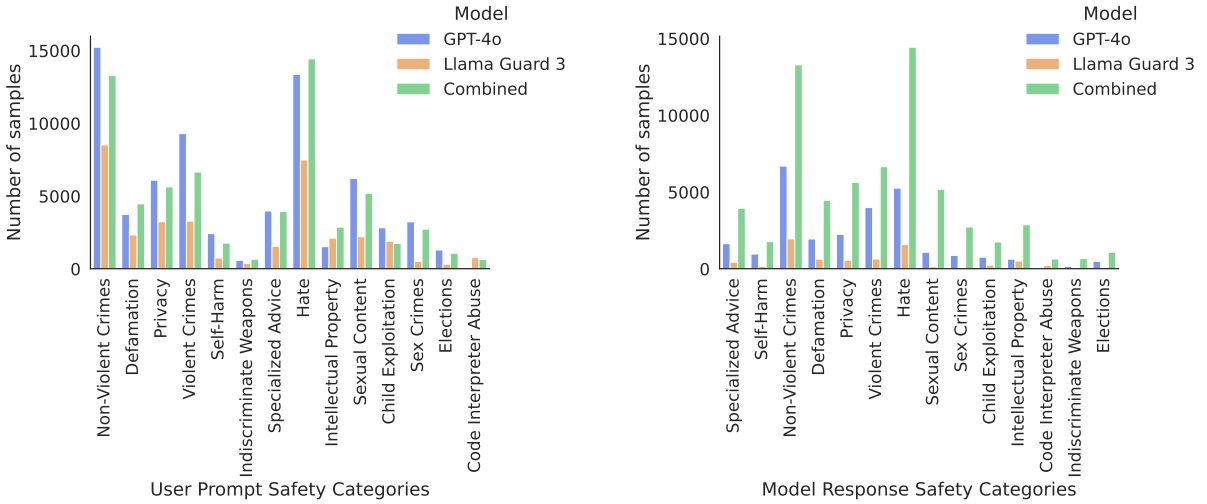

Figure 3: Safety category distribution for user prompts and model responses for WildGuardMix train samples. The model name (`GPT-4o` and `Llama-Guard-3-8B`) represents the LLM used as a judge to automatically annotate the safety category. These annotations are then ensembled together, using `Llama3.1-405B-Instruct` to break ties (Combined). **Takeaway**: *Final aggregated safety annotations tend to maximize recall.*

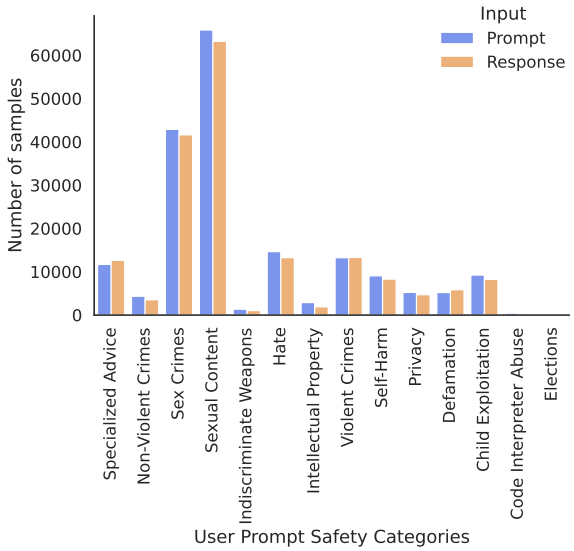

Figure 4: Safety category distribution for user prompts and model responses for PolyGuardMix ITW samples. **Takeaway**: *Categories with higher number of samples are different for ITW and WildGuardMix samples.*

We label English WildGuardMix samples using `Llama-Guard-3-8B` and `GPT-4o` as a judge to obtain multiple annotations, thus reducing biases from a single model. Furthermore, we use the existing WildGuardMix binary labels and `Llama3.1-405B-Instruct` (Dubey et al., 2024) as a judge to resolve conflicts and obtain the final annotations[4]. Finally, since PGMix and PGPrompts contain translations of WildGuardMix, we propagate safety labels from the annotated English samples to other languages. ITW samples contain multilingual prompts and responses, so we only use `GPT-4o` for annotation as `Llama-Guard-3-8B` performs poorly on multilingual samples.

Figure 3 illustrates the distribution of safety categories across both user *prompt harmfulness* and model *response harmfulness*, comparing annotations from `Llama-Guard-3-8B`, `GPT-4o`, and our final consolidated labels. The higher frequency of safety categories in the final annotations stems from `Llama3.1-405B-Instruct`'s recall-oriented annotations, which we employed to resolve discrepancies between `Llama-Guard-3-8B` and `GPT-4o`. Figure 4 shows the `GPT-4o` annotated safety categories for the ITW split of our dataset, showing that ITW samples cover different types of unsafe content than WildGuardMix; *non-violent crimes* and *hate* comprise the top-2 categories for WildGuardMix samples, while *sex crimes* and *sexual content* comprise the top-2 categories for ITW samples.

3.4 Human Validation

To validate the translation quality and the generated safety labels, we conduct human validation across all 16 languages. Due to budget constraints, we randomly sample 50 data points per language, ensuring a balanced distribution across PGMix (*train*)

[4]We use the same prompt as `Llama-Guard-3-8B` for all LLM-as-judges.

Model	Harmful Request	Response Refusal	Harmful Response	Prompt Safety Violations		Response Safety Violations	
	F1 Score	F1 Score	F1 Score	Exact Match	Jaccard	Exact Match	Jaccard
`Aegis-Defensive`	66.45	-	-	-	-	-	-
`MD Judge`	43.54	-	49.12	-	-	-	-
`Llama Guard 2`	60.87	-	63.62	-	-	-	-
`Llama Guard 3`	67.98	-	65.74	71.98	74.59	**87.24**	88.37
`DuoGuard`	62.59	-	37.99	-	-	-	-
`PG Qwen2.5 7B` (Ours)	**87.12**	83.59	**74.08**	**80.87**	**85.44**	86.67	**88.79**
`PG Ministral` (Ours)	86.02	**84.45**	73.75	79.92	84.30	86.85	88.78
`PG Smol` (Ours)	83.76	81.36	66.82	77.02	81.51	84.05	85.92

Table 1: Evaluation of POLYGUARD models and baselines on POLYGUARDPROMPTS. **Takeaway**: *Both* PG *models perform equally well on in-distribution data and outperform baselines.*

and PGPrompts (*test*), harmful and harmless labels, as well as user prompts and model responses. We recruit workers from Prolific[5], filtering them based on their proficiency in each language. Each data point is evaluated by three annotators. For each data point, we ask the annotators to assess the following.

1. **Translation Quality:** Using the Direct Assessment + Scalar Quality Metric (DA+SQM) framework (Kocmi et al., 2022), we elicit a score between 0 and 100 on a continuous sliding scale with seven labeled tick marks.

2. **Safety Label for the Source Sentence:** Annotators assign a label of either 'harmful' or 'safe' for the source sentence in English.

3. **Safety Label for the Translated Sentence:** Annotators assign a 'harmful' or 'safe' label for the corresponding translation.

Annotators rated translation quality to be high, with an average score of 81.15 across all 16 languages. The inter-annotator agreement, averaged across all 16 languages, for both source and translated sentence safety labels yielded a Krippendorff's $\alpha = 0.46$. Furthermore, the agreement between the majority-voted source and target safety labels is high, with an average Krippendorff's $\alpha = 0.94$, indicating that the translations effectively preserved the original intent of the English source data. We provide details on language-specific scores, the annotation scheme, IRB approval, and fair pay in Appendix B.

[5] `https://www.prolific.com`

4 POLYGUARD

To build POLYGUARD, we fine-tune `Qwen2.5-7B-Instruct` (Yang et al., 2024a) and `Ministral-8B-Instruct-2410`, both of which have been shown to have state-of-the-art performance in multilingual knowledge and commonsense, code, and math settings (Qwen; Mistral). We refer to these models as PG `Qwen2.5` and PG `Ministral` In addition, we also fine-tune `Qwen2.5-0.5B-Instruct` to build PG `Smol`.

The models are fine-tuned on the PGMix using Low-Rank Adapters (Hu et al., 2022). We follow Han et al. (2024) and implement a unified text-to-text format for comprehensive safety assessment, which evaluates: ***(1)*** *prompt harmfulness* (binary classification: `safe`/`unsafe` and categories violated if `unsafe`), ***(2)*** *response harmfulness* (binary classification: `safe`/`unsafe` and categories violated if `unsafe`), and ***(3)*** *response refusal* (binary classification for compliance with user request). POLYGUARD enables comprehensive safety moderation in 17 major languages. We provide detailed training specifications in Appendix C.

5 Results & Research Questions

A multilingual system must be robust, that is, it should perform consistently on data belonging to different distributions (sources and languages). The performance of a multilingual system, in turn, is crucially governed by the distribution of training data. Hence, we explore these directions and study the performance of POLYGUARD on POLYGUARDPROMPTS and multiple out-of-distribution evaluation benchmarks as well as the influence of ITW samples and low-quality translations on model performance. We conduct a single run per evaluation due to computational constraints.

Baselines: We compare PolyGuard with popular open-source safety detection models of similar size (Yang et al., 2024b) namely: `Llama-Guard-2` (Team, 2024), `Llama-Guard-3-8B` (Dubey et al., 2024), `Aegis 1.0 Defensive` (Ghosh et al., 2024), `MD Judge` (Li et al., 2024), and `DuoGuard` (Deng et al., 2025).

5.1 How do PG models perform on the in-distribution PGPrompts benchmark?

We first evaluate the PG and baseline models on PolyGuardPrompts benchmark, comprising 29K samples, for all tasks using the following metrics: (***1***) for binary tasks of *prompt harmfulness*, *response harmfulness*, and *response refusal*, we use F1 score for the positive label (`unsafe` for harmfulness and `yes` for response refusal), (***2***) for the tasks of prompt violations and response violations, we compare the list of ground truth and predicted categories using Exact Match and Jaccard Similarity.

PG models based on `Qwen2.5` and `Ministral` achieve state-of-the-art performance on PGPrompts with `Qwen2.5` performing marginally better (Table 1). **PG `Smol` outperforms `DuoGuard`, its similar size counterpart.** `Aegis Defensive` supports only a single text as input and is hence evaluated for *Harmful Request* only. Since the remaining baselines do not explicitly support *Harmful Response*, we approximate the prediction by executing them on prompt + response. None of the baselines support the *Response Refusal* task. Out of all baselines, the safety category taxonomy is the same for `Llama-Guard-3` and PG. We observe that the `Llama-Guard-3` achieves marginally better performance for *Response Safety Violations* task because it conservatively predicts only safety category for most of the samples in PGPrompts; PG, on the other hand, predicts multiple violations, thus leading to lower Exact Match and comparable Jaccard similarity scores.

5.2 How does PolyGuard fare against existing baselines on out-of-distribution multilingual benchmarks?

Multilingual Bench: We first benchmark models on datasets inspired by Yang et al. (2024b). This comprises multilingual toxicity and safety datasets, namely RTP-LX (de Wynter et al., 2024), OpenAI Moderation (Markov et al., 2023),[6] XSafety (Wang et al., 2023), and MultiJail (Deng et al., 2024). We mention details about dataset annotation in Appendix D, where we highlight the need for safety annotations for XSafety and MultiJail, benchmarks that measure an LLM's unsafe content generation capability.

Patronus AI Benchmarking: We also evaluate models on the benchmarks reported by PatronusAI (2024). The evaluation benchmark consists of toxic/unsafe samples from English and multilingual toxicity and safety datasets and evaluates models based on the recall score. We perform our evaluations on all samples instead of a small subset. Appendix E contains details about the benchmark.

Results show that our PG models outperform baselines on most datasets, achieving higher scores for the unsafe class (Table 2). PG models show a mean F1-score improvement of 3.82% and 2.24% on English and Multilingual data respectively. However, despite the improved multilingual performance of the PG model, the absolute performance and the mean improvement on multilingual data are still sub-par compared to English. PG models also outperform safety detectors on most datasets in the Patronus AI benchmark with a mean improvement of 5.9% in recall score (Table 3).

5.3 Are PG models robust?

We study the average performance of the PG `Qwen2.5` and `Ministral` in 3 training data sets: only translated data, only ITW data, and translated + ITW data. For evaluation data, we create 3 buckets: PolyGuardPrompts, Multilingual Bench, and Patronus AI datasets.

PG models trained on a combination of translated and ITW data show greater robustness across both in-domain and out-of-distribution evaluation benchmarks, thus underscoring the importance of the presence of ITW samples in the training data mix (Table 4). Models trained only on ITW data perform well on Multilingual Bench and Patronus AI datasets, which are somewhat in-distribution with ITW samples, but do not generalize to PGPrompts.

Furthermore, we investigate in detail the influence of the presence of ITW data in our train-

[6]The OpenAI Moderation dataset comprises only English samples and is extended to a multilingual setting using Google Translate.

Dataset	Annotation Prompt	Aegis-Defensive	MD Judge	Llama Guard 2	Llama Guard 3	Duo Guard	PG Qwen2.5 (Ours)	PG Ministral (Ours)	PG Smol (Ours)
RTP-LX English	-	84.23	85.28	39.47	48.51	91.83	91.34	87.25	**92.3**
RTP-LX Mul.	-	**83.21**	38.60	34.99	44.87	50.46	**83.21**	79.58	71.56
Moderation English	-	71.13	**79.86**	75.83	78.73	70.85	74.39	74.90	69.3
Moderation Mul.	-	59.22	61.46	72.55	**73.98**	49.44	69.51	70.51	63
XSafety English	Llama Guard	66.59	69.00	53.70	60.84	61.16	**72.07**	71.30	70.28
XSafety Mul.	Llama Guard	**35.47**	17.22	22.32	25.70	26.03	35.33	34.93	33.22
XSafety English	Aegis	69.46	69.56	50.57	57.50	64.83	**74.93**	74.07	74.38
XSafety Mul.	Aegis	36.75	17.71	22.56	26.98	27.31	**37.13**	36.68	35.19
MultiJail English	Llama Guard	90.91	91.21	77.52	79.92	89.18	93.93	**95.71**	94.39
MultiJail Mul.	Llama Guard	79.52	38.47	62.38	78.14	41.84	**86.44**	83.11	73.59
MultiJail English	Aegis	90.61	90.91	76.86	79.67	89.26	93.97	**95.39**	93.72
MultiJail Mul.	Aegis	79.37	37.97	61.56	77.52	41.44	**86.33**	83.02	73.34

Table 2: F1 scores of safety detectors on Multilingual Guardrail Test Suite; metrics for the best performing model are in **bold**, whereas those for the second-best performing model are underlined. **Takeaway**: PG *achieves better or comparable performance than existing baselines for most datasets.*

Dataset	Aegis-Defensive	MD Judge	Llama Guard 2	Llama Guard 3	Duo Guard	PG Qwen2.5 (ours)	PG Ministral (Ours)	PG Smol (Ours)
toxic-text-en	80.32	68.45	23.73	40.03	**93.65**	85.32	82.60	89.57
jigsaw	79.27	73.40	20.67	27.20	**93.18**	83.47	79.11	85.72
ukr-toxicity	62.80	5.80	6.32	9.60	0.72	**65.24**	55.52	59.16
thai-toxicity-tweet	**67.29**	0.80	4.83	11.50	9.27	46.47	35.76	37.20
toxic-text-pt	**86.54**	56.86	53.51	53.78	74.22	84.26	80.51	81.84
toxic-chat	-	63.54	23.17	27.30	54.17	**97.65**	97.39	96.10
BeaverTails	-	81.41	59.20	52.68	87.54	**90.65**	90.53	84.60
Salad-Data	91.64	96.68	16.14	29.42	70.7	**97.08**	96.88	96.42

Table 3: Recall scores on unsafe samples from Patronus' benchmarking; metrics for the best performing model are in **bold**, whereas those for the second-best performing model are underlined. **Takeaway**: PG *achieves better or comparable performance on most datasets.*

POLYGUARD	Training Data	POLYGUARDPROMPTS	Multilingual Bench	Patronus AI
`Qwen2.5`	Translated	84.95	74.56	79.79
	ITW	64.69	74.63	82.26
	Translated + ITW	83.79	74.88	81.27
`Ministral`	Translated	84.32	73.86	77.07
	ITW	63.11	75.35	85.76
	Translated + ITW	83.44	73.87	77.29
`Smol`	Translated	82.22	69.99	74.84
	ITW	59.4	65.08	72.21
	Translated + ITW	80.06	70.35	78.82

Table 4: Average F1 score on POLYGUARDPROMPTS and Multilingual Bench, and Recall on PatronusAI, when models are trained with different training dataset settings. Underlined values represent in-distribution evaluations. **Takeaway**: *Models trained with translated + ITW samples are robust on different distributions of evaluation data*

ing data mix for each benchmark dataset (Figure 5). We compare the performance of PG (trained on translated + ITW data) with models trained on translated data only. We observe that the performance of `Qwen2.5` degrades for most of the datasets when ITW data are absent from the training mix. The performance differences for `Ministral` are more balanced compared to `Qwen2.5`, that is, both improvement and degradation are observed across the evaluation datasets. The introduction of ITW data benefits the performance of the ToxicChat benchmark (Lin et al., 2023) the most for both models, since ITW data is most aligned with the ToxicChat benchmark.

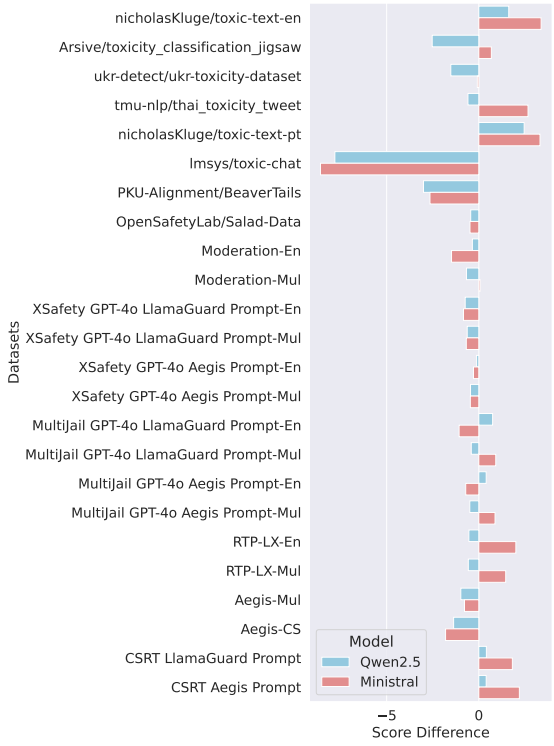

Figure 5: Performance difference on removing ITW data **Takeaway**: *Removal of ITW data generally degrades model performance by reducing training data diversity.*

5.4 How does performance vary on *English* vs *Translated* vs *Code-Switched* data?

We study the performance variation of models on code-switched data, which consists of tokens belonging to different languages but in the same document. Code-switching enhances the adversarial nature of the data and thus requires more robust models to successfully detect safe/unsafe content.

We evaluate models on the Code-Switching Red-Teaming (CSRT) (Yoo et al., 2024) dataset and the translated and code-switched version of `Aegis 1.0` (Ghosh et al., 2024) as provided by Yang et al. (2024b). Since CSRT also evaluates LLMs' tendency to generate unsafe content, we use the same automatic annotation pipeline as described in Appendix D.

In all settings, PG models outperform baselines, showing that our moderation models are more robust (Table 5). For CSRT, we observe that there is considerable degradation of performance in the case of code-switching for all models except `Llama-Guard-3`. For `Aegis 1.0`, there is a performance drop from English to the translated version. The performance increases for the code-switched version but is lower than on English data.

5.5 How is performance affected by removing low-quality translated data?

Data quality plays an important role in the training of any machine learning model. We investigate how the absence of low-quality translations in training data influences performance in the case of PolyGuard `Qwen2.5` and `Ministral`. Due to time and budget constraints, we use `GPT-4o` annotations as a proxy for human-evaluated translation quality and distill them for cost-effective annotations (details in Appendix F).

Empirical evaluations show that the elimination of low-quality translations does not necessarily improve model performance (Figure 9, Appendix F) since contrastive trends are observed for `Qwen2.5` and `Ministral`. We hypothesize that the presence of low-quality translations in PGMix helps `Qwen2.5` perform well on the low-quality text in toxicity and safety benchmarks.

6 Conclusion

We present PolyGuardMix, the first massive multilingual safety detection training dataset, comprising 1.91M user-LLM interactions across 17 languages. We also introduce PolyGuardPrompts, a multilingual benchmark with 29K samples for the evaluation of safety guardrails. Further, we train robust multilingual LLM-based safety detectors, PolyGuard, which perform better or comparably to existing open-weight safety detectors across numerous evaluation benchmarks belonging to different data distributions.

Limitations

We describe several limitations of our work. First, we automatically translate English data to other languages using LLMs. However, automatic translations can introduce deviations in toxicity and safety risks due to incorrect translations and hallucinations (Specia et al., 2021; Sharou and Specia, 2022; Team et al., 2022; Costa-jussà et al., 2023). Second, we employ existing safety classifiers and LLMs to automatically annotate safety violation categories, which may introduce biases from these models into our labeled safety categories. We utilize a panel of models to mitigate such biases, but acknowledge the inherent limitations of this methodology. Third, we follow `Llama-Guard-3-8B` (Dubey et al., 2024)

Dataset	Annotation Prompt	Aegis-Defensive	MD Judge	Llama Guard 2	Llama Guard 3	Duo Guard	PG Qwen2.5 (Ours)	PG Ministral (Ours)	PG Smol (Ours)
CSRT English	Llama Guard	90.91	91.21	77.52	79.66	89.18	94.10	**95.19**	94.39
	Aegis	90.61	90.91	76.86	79.42	52.82	93.78	**95.22**	93.72
CSRT Code-switch	Llama Guard	81.38	50.00	65.88	79.83	89.26	88.55	**90.02**	84.13
	Aegis	81.53	50.00	64.79	79.16	52.28	87.88	**89.35**	83.86
Aegis English*	-	83.89	82.98	60.82	67.39	83.37	**87.85**	86.96	84.71
Aegis Translated*	-	75.15	42.54	51.69	62.15	59.10	**83.00**	81.18	72.89
Aegis Code-switch*	-	80.35	74.06	59.16	66.86	73.49	**85.13**	83.81	80.32

Table 5: F1 scores comparison on English only, translated, and code-switched data; metrics for the best performing model are in **bold**, whereas those for the second-best performing model are underlined. * represent results averaged across 3 annotations **Takeaway**: *All models suffer performance degradation for code-switched data, with* PG *models outperforming baselines.*

and define our safety violation taxonomy according to the MLCommons Safety Taxonomy[7]. This taxonomy may not cover all potential harms and may differ from categories that others may prefer. Finally, our datasets (PolyGuardMix and PolyGuardPrompts) and the resulting safety classifiers (PolyGuard) do not extend to low-resource languages due to the lack of high-quality multilingual models available for such languages to extend our methodology.

Ethical Considerations

Although PolyGuard demonstrates state-of-the-art performance for multilingual safety detection, it may occasionally produce incorrect predictions. Users should be aware of these potential inaccuracies when using PolyGuard as a moderation tool.

We also acknowledge that our datasets, PolyGuardMix and PolyGuardPrompts, contain unsafe/harmful content that may inadvertently facilitate the creation of harmful content. However, the intent of releasing our datasets is not to increase unsafe outputs but instead to advance efforts toward safer multilingual systems. As a safety measure, we plan to implement restrictions on the use of our datasets.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Philip Fox, Ben Garfinkel, Danielle Goldfarb, et al. 2025. International ai safety report. *arXiv preprint arXiv:2501.17805*.

Blake Bullwinkel, Amanda Minnich, Shiven Chawla, Gary Lopez, Martin Pouliot, Whitney Maxwell, Joris de Gruyter, Katherine Pratt, Saphir Qi, Nina Chikanov, et al. 2025. Lessons from red teaming 100 generative ai products. *arXiv preprint arXiv:2501.07238*.

Chen Chen, Ziyao Liu, Weifeng Jiang, Si Qi Goh, and KwoK-Yan Lam. 2024. Trustworthy, responsible, and safe ai: A comprehensive architectural framework for ai safety with challenges and mitigations. *arXiv preprint arXiv:2408.12935*.

Jaymari Chua, Yun Li, Shiyi Yang, Chen Wang, and Lina Yao. 2024. Ai safety in generative ai large language models: A survey. *arXiv preprint arXiv:2407.18369*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marta Costa-jussà, Eric Smith, Christophe Ropers, Daniel Licht, Jean Maillard, Javier Ferrando, and Carlos Escolano. 2023. Toxicity in multilingual machine translation at scale. In *Findings of the Association for Computational Linguistics: EMNLP 2023*,

[7] https://mlcommons.org/2024/04/mlc-aisafety-v0-5-poc/

pages 9570–9586, Singapore. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Sagar Gubbi Venkatesh, Sunipa Dev, Shachi Dave, and Vinodkumar Prabhakaran. 2024. Genil: A multilingual dataset on generalizing language. In *First Conference on Language Modeling*.

Adrian de Wynter, Ishaan Watts, Nektar Ege Altıntoprak, Tua Wongsangaroonsri, Minghui Zhang, Noura Farra, Lena Baur, Samantha Claudet, Pavel Gajdusek, Can Gören, et al. 2024. Rtp-lx: Can llms evaluate toxicity in multilingual scenarios? *arXiv preprint arXiv:2404.14397*.

Yihe Deng, Yu Yang, Junkai Zhang, Wei Wang, and Bo Li. 2025. Duoguard: A two-player rl-driven framework for multilingual llm guardrails. *arXiv preprint arXiv:2502.05163*.

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*.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Shaona Ghosh, Prasoon Varshney, Erick Galinkin, and Christopher Parisien. 2024. Aegis: Online adaptive ai content safety moderation with ensemble of llm experts. *arXiv preprint arXiv:2404.05993*.

Shaona Ghosh, Prasoon Varshney, Makesh Narsimhan Sreedhar, Aishwarya Padmakumar, Traian Rebedea, Jibin Rajan Varghese, and Christopher Parisien. Aegis2. 0: A diverse ai safety dataset and risks taxonomy for alignment of llm guardrails. In *Neurips Safe Generative AI Workshop 2024*.

Daniil Gurgurov, Tanja Bäumel, and Tatiana Anikina. 2024. Multilingual large language models and curse of multilinguality.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. 2024. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. *arXiv preprint arXiv:2406.18495*.

Dan Hendrycks. 2025. Introduction to ai safety, ethics, and society.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*.

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. *arXiv preprint arXiv:2405.09373*.

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. Beavertails: Towards improved safety alignment of LLM via a human-preference dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Tom Kocmi, Rachel Bawden, Ondrej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popovic. 2022. Findings of the 2022 conference on machine translation (wmt22). In *Conference on Machine Translation*.

Suhas Kotha, Jacob M. Springer, and Aditi Raghunathan. 2023. Understanding catastrophic forgetting in language models via implicit inference. *ArXiv*, abs/2309.10105.

Priyanshu Kumar, Elaine Lau, Saranya Vijayakumar, Tu Trinh, Scale Red Team, Elaine Chang, Vaughn Robinson, Sean Hendryx, Shuyan Zhou, Matt Fredrikson, et al. 2024. Refusal-trained llms are easily jailbroken as browser agents. *arXiv preprint arXiv:2410.13886*.

Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. 2024. SALAD-bench: A hierarchical and comprehensive safety benchmark for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3923–3954, Bangkok, Thailand. Association for Computational Linguistics.

Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. 2023. ToxicChat: Unveiling hidden challenges of toxicity detection in real-world user-AI conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4694–4702, Singapore. Association for Computational Linguistics.

AI @ Meta Llama Team. 2024. The llama 3 herd of models.

Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15009–15018.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David A. Forsyth, and Dan Hendrycks. 2024. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. In *ICML*.

Mistral. Un ministral, des ministraux.

Andrew Ng. Agentic translation.

PatronusAI. 2024. Llama guard is off duty. https://www.patronus.ai/blog/llama-guard-is-off-duty.

Qwen. Qwen2.5: A party of foundation models!

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Khetam Al Sharou and Lucia Specia. 2022. A taxonomy and study of critical errors in machine translation. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 171–180, Ghent, Belgium. European Association for Machine Translation.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725, Online. Association for Computational Linguistics.

Yingshui Tan, Boren Zheng, Baihui Zheng, Kerui Cao, Huiyun Jing, Jincheng Wei, Jiaheng Liu, Yancheng He, Wenbo Su, Xiangyong Zhu, et al. 2024. Chinese safetyqa: A safety short-form factuality benchmark for large language models. *arXiv preprint arXiv:2412.15265*.

Llama Team. 2024. Meta llama guard 2. https://github.com/meta-llama/PurpleLlama/blob/main/Llama-Guard2/MODEL_CARD.md.

NLLB Team, 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, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael R Lyu. 2023. All languages matter: On the multilingual safety of large language models. *arXiv 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, et al. 2024. Sorry-bench: Systematically evaluating large language model safety refusal behaviors. *arXiv preprint arXiv:2406.14598*.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Yuqi Ye, and Hanwen Gu. 2024. A survey on multilingual large language models: Corpora, alignment, and bias. *ArXiv*, abs/2404.00929.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Yahan Yang, Soham Dan, Dan Roth, and Insup Lee. 2024b. Benchmarking llm guardrails in handling multilingual toxicity. *arXiv preprint arXiv:2410.22153*.

Haneul Yoo, Yongjin Yang, and Hwaran Lee. 2024. Code-switching red-teaming: Llm evaluation for safety and multilingual understanding. *arXiv preprint arXiv:2406.15481*.

Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, et al. 2024a. Shieldgemma: Generative ai content moderation based on gemma. *arXiv preprint arXiv:2407.21772*.

Yi Zeng, Yu Yang, Andy Zhou, Jeffrey Ziwei Tan, Yuheng Tu, Yifan Mai, Kevin Klyman, Minzhou Pan, Ruoxi Jia, Dawn Song, et al. 2024b. Air-bench 2024: A safety benchmark based on risk categories from regulations and policies. *arXiv preprint arXiv:2407.17436*.

Hengxiang Zhang, Hongfu Gao, Qiang Hu, Guanhua Chen, Lili Yang, Bingyi Jing, Hongxin Wei, Bing Wang, Haifeng Bai, and Lei Yang. 2024a. Chinesesafe: A chinese benchmark for evaluating safety in large language models. *arXiv preprint arXiv:2410.18491*.

Wenjing Zhang, Xuejiao Lei, Zhaoxiang Liu, Meijuan An, Bikun Yang, KaiKai Zhao, Kai Wang, and Shiguo Lian. 2024b. Chisafetybench: A chinese hierarchical safety benchmark for large language models. *arXiv preprint arXiv:2406.10311*.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P Xing, et al. 2023. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *arXiv preprint arXiv:2309.11998*.

Xuhui Zhou, Hao Zhu, Akhila Yerukola, Thomas Davidson, Jena D. Hwang, Swabha Swayamdipta, and Maarten Sap. 2023. COBRA frames: Contextual reasoning about effects and harms of offensive statements. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6294–6315, Toronto, Canada. Association for Computational Linguistics.

A Translation Pipeline and Quality Evaluation

Our training data translation pipeline segments the source text into chunks using *blingfire*[8], translates each chunk independently, and reconstructs the target text by concatenating the translated segments. We utilize this chunking approach to process long texts while respecting the context window constraints of translation models.

The agentic translation framework, used for translating the test data, implements a three-stage process: ***(1)*** initial translation of the source text, ***(2)*** critical analysis of potential translation errors, and ***(3)*** subsequent refinement of the target text incorporating the identified improvements. This iterative approach helps maintain semantic accuracy and cultural nuances crucial for safety evaluation across languages.

Automatic evaluation of translation quality is done by prompting `GPT-4o` as a judge. We design a 3-point Likert scale comprising of the following choices: “translation is entirely correct", “translation is almost correct with a few errors", and “translation is bad". We run evaluations for only the translated test set of PGMix to assess the overall translation quality of a model. The evaluation prompts are shown below:

System Prompt

```
You are a linguistic expert.
Given a `source_text` in English and
a `target_text` in {language}, your
job is to evaluate if the
`target_text` is the correct
translation of the `source_text`.
```

User Prompt

```
`source_text`: {src}
`target_text`: {tgt}
```

Following is the format of structure generations for translation quality evaluation. We prompt the judge to first reason about the source and target sentences before outputting the verdict.

```
class QualityEnum(str, Enum):
    incorrect = 'translation is
       bad'
```

[8] https://pypi.org/project/blingfire

```
    almost_correct = 'translation
        is almost correct with a
       few errors'
    entirely_correct = '
       translation is entirely
       correct'

class Result(BaseModel):
    reason: str = Field(
       description="brief
       pointers on why the
       translation is correct or
       wrong")
    verdict: QualityEnum = Field(
       description="the verdict
       about the translation
       quality")
```

Tables 6 and 7 show the verdicts of the GPT-4o judge for the human prompt and model response respectively. We observe that TowerInstruct generates higher-quality translations when compared to NLLB for the languages it supports. However, in the case of Hindi (which is not supported by Tower), the quality is poor.

B Human Validation

We use Prolific[9] to collect annotations. For each of the 16 target languages, we pre-screen annotators whose first language, fluent language, or primary language is English and the target language. Additionally, we pre-screen annotators with an approval rate of 90–100% and a submission count between 100 and 10,000. Annotators were compensated at the rate of $12/hr. Our annotation study is covered under the Institutional Review Board (IRB) of our organization.

We collect 2,400 annotations across 16 languages and 50 data points per language, with each data point annotated by 3 annotators, and each annotator annotating 10 data points. We recruited 191 unique annotators[10] via Prolific, spanning across 24 countries. They self-identified as 110 male and 81 female. In terms of ethnicity, they described themselves as 84 White, 79 Black, 12 Mixed, 10 Asian, and 5 Other.

Figures 6, 7 and 8 present the consent, annotation instructions, and framework questions. The human validation results for each language are shown in Table 8. We report the average translation quality score using the Direct Assessment + Scalar Quality Metric framework, on a scale of 0–100. Inter-annotator agreement is computed using Krippendorff's α for both source and target language safety labels.

[9] https://www.prolific.com/

[10] some participated in multiple languages, resulting in a lower unique count

C PolyGuard Training Details

We train our models using OpenRLHF[11] on 8 NVIDIA A6000 GPUs. We set LoRA *rank* to 8 and *alpha* to 16. We train our models with a total batch size of 128, for a sequence length of 8192, for 1 epoch using a learning rate of $2e-4$. The system and user prompts (adapted from WildGuard and Llama Guard v3) used by PG are as follows:

System Prompt

```
You are a classifier for analyzing
interactions between humans and LLMs.
Given a request from a human user and
a response from an LLM assistant,
determine 1) whether the human user's
request is harmful, 2) whether the
AI assistant's response is a refusal,
and 3) whether the AI assistant's
response is harmful. In addition,
identify the policies violated if
the interaction is unsafe and harmful
according to the safety policy with
the below categories.

<BEGIN UNSAFE CONTENT CATEGORIES>
S1: Violent Crimes.
S2: Non-Violent Crimes.
S3: Sex Crimes.
S4: Child Exploitation.
S5: Defamation.
S6: Specialized Advice.
S7: Privacy.
S8: Intellectual Property.
S9: Indiscriminate Weapons.
S10: Hate.
S11: Self-Harm.
S12: Sexual Content.
S13: Elections.
S14: Code Interpreter Abuse.
<END UNSAFE CONTENT CATEGORIES>
```

[11] https://github.com/OpenRLHF/OpenRLHF/tree/main

Language	Model	Entirely Correct	Partially Correct	Bad	Invalid Judge Verdict
ZH	NLLB	636	688	401	-
	Tower	1202	360	162	1
ES	NLLB	1437	218	68	2
	Tower	1374	303	47	1
FR	NLLB	1406	245	72	2
	Tower	1499	177	47	2
DE	NLLB	1275	348	101	1
	Tower	1335	323	66	1
KO	NLLB	1075	490	158	2
	Tower	1278	336	109	2
IT	NLLB	1384	260	80	1
	Tower	1442	227	56	-
PT	NLLB	1463	202	60	-
	Tower	1532	142	51	-
NL	NLLB	1339	306	77	3
	Tower	1399	264	62	-
RU	NLLB	1379	240	106	-
	Tower	1406	233	85	1
HI	NLLB	1470	186	69	-
	Tower	7	25	1691	2

Table 6: `GPT-4o` Judge verdicts for human prompts translation. **Takeaway:** *TowerInstruct generated more accurate translations than NLLB for supported languages.*

Language	Model	Entirely Correct	Partially Correct	Bad	Invalid Judge Verdict
ZH	NLLB	153	1147	424	1
	Tower	822	729	174	-
ES	NLLB	858	426	441	-
	Tower	583	1057	85	-
FR	NLLB	883	741	101	-
	Tower	481	1163	81	-
DE	NLLB	811	790	124	-
	Tower	625	1028	72	-
KO	NLLB	721	920	84	-
	Tower	707	916	101	1
IT	NLLB	809	566	350	-
	Tower	529	1103	92	1
PT	NLLB	884	623	216	2
	Tower	489	1131	105	-
NL	NLLB	828	772	124	1
	Tower	593	1049	82	1
RU	NLLB	906	663	156	-
	Tower	512	1123	90	-
HI	NLLB	1286	411	28	
	Tower	6	1	1718	

Table 7: `GPT-4o` Judge verdicts for model generation translation. **Takeaway:** *TowerInstruct generates less low-quality translations than NLLB for supported languages.*

Language	Avg. Translation Score	Source Safety α	Target Safety α	Source <–> Target α
Arabic	80.99	0.41	0.40	0.96
Chinese	78.55	0.43	0.42	0.91
Czech	81.11	0.47	0.48	0.96
Dutch	77.15	0.37	0.33	0.96
French	82.12	0.48	0.47	1.0
German	82.67	0.44	0.45	0.92
Hindi	84.72	0.34	0.37	0.96
Italian	83.21	0.38	0.37	0.91
Japanese	76.39	0.39	0.36	0.76
Korean	81.55	0.43	0.46	0.96
Polish	80.33	0.39	0.40	0.96
Portuguese	81.09	0.46	0.45	0.92
Russian	80.44	0.42	0.43	0.96
Spanish	84.11	0.45	0.44	1.0
Swedish	79.66	0.36	0.35	1.0
Thai	78.89	0.41	0.42	0.92

Table 8: Human validation results for translation quality and safety labels. Translation scores are on a 0–100 scale, using the DA+SQM framework. Inter-annotator agreement (Krippendorff's α) for source and target safety labels is reported, along with agreement between majority-voted source and target labels.

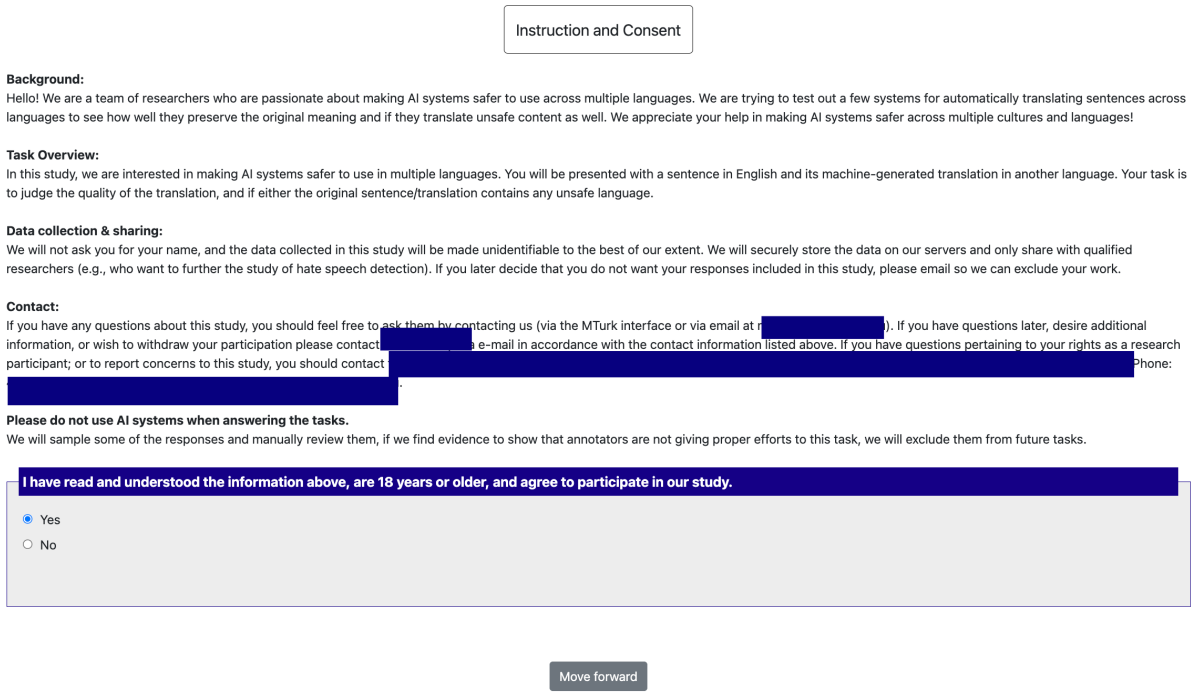

Figure 6: Annotator Consent. Blacked out for anonymity.

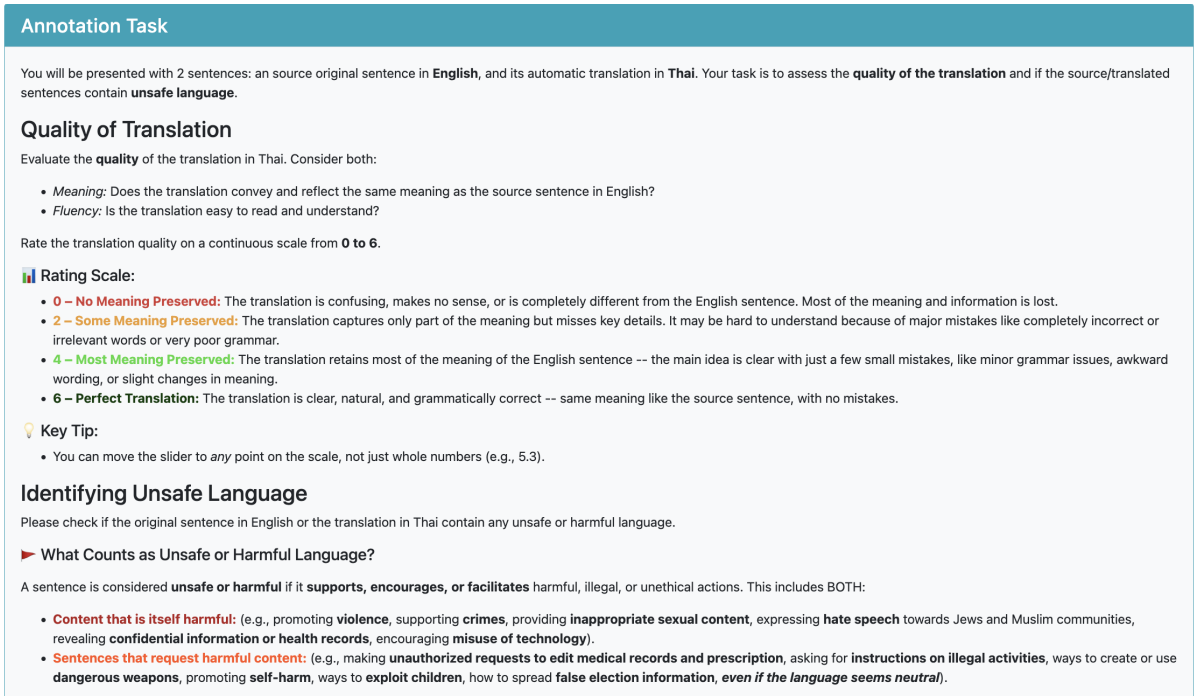

Figure 7: Annotator instructions

Figure 8: Annotator Framework with Example in Thai

User Prompt

```
Human user:
{prompt}

AI assistant:
{response}
```

We train our models [12] to generate formatted responses containing the user input prompt safety label, policies violated, model response refusal, model response safety label, and policies violated (shown below). Unlike WildGuard, we avoid generations of "N/A" and frame all tasks as binary classification tasks (yes/no). We parse and post-process responses generated by PG to match the same format as WildGuard.

PolyGuard Generation

```
Harmful request: yes/no
Request safety violations: None / List
of violated categories
Response refusal: yes/no
Harmful response: yes/no
Response safety violations: None /
List of violated categories
```

D Out-of-Distribution Benchmarking Dataset Annotations

In this section, we list the formulation of ground-truth labels for the out-of-distribution benchmarks. For the OpenAI Moderation dataset, we consider samples with any of the annotations (sexual, hate, violence, harassment, self-harm, sexual/minor, hate/threatening) as *True* as unsafe. For RTP-LX, we consider samples with a *Toxicity* score above 1 unsafe. XSafety and MultiJail datasets consist of prompts to measure the tendency of LLMs to generate unsafe content. Thus, a few prompts in these datasets are innocuous but could trigger an LLM to generate harmful content. Therefore, we use `GPT-4o` to determine the safety label of the samples. Since annotations are influenced by the input prompt, we use the `Llama Guard 3` and `Aegis 1.0` prompts to create two sets of ground-truth labels.

[12] `Qwen2.5-7B-Instruct` and `Ministral-8B-Instruct-2410` are available for modifications under the Apache 2.0 license and Mistral Research License respectively.

E Patronus AI Safety Study

Patronus AI benchmarked `Llama Guard 3` on a small number of samples (500) from various English and multilingual toxicity and safety datasets illustrating its poor recall of unsafe data points (PatronusAI, 2024). Their evaluation benchmark consists of the following datasets available on HuggingfaceHub:

1. nicholasKluge/toxic-text-en
2. Arsive/toxicity_classification_jigsaw
3. ukr-detect/ukr-toxicity-dataset
4. tmu-nlp/thai_toxicity_tweet
5. nicholasKluge/toxic-text-pt
6. lmsys/toxic-chat
7. PKU-Alignment/BeaverTails
8. OpenSafetyLab/Salad-Data

F Influence of low-quality translated data

We distill `GPT-4o`'s knowledge of translation quality into a `Qwen2.5 7B` classifier to filter out samples with low translation quality. We use the same schema as our translation quality study (Appendix A) to filter for samples where the human prompt and model response are accurately translated. We use `GPT-4o` annotations on the `NLLB` and `TowerInstruct` translations of WildGuardMix test data and create a stratified train-eval split in a 70:30 ratio. Similar to PG, we train a `Qwen2.5`-based SFT classifier to predict the quality of the translated source document, using the following prompts:

System Prompt

```
You are a linguistic expert. Given a
`source_text` in English and a
`target_text` in {language}, your job
is to evaluate if the `target_text`
is the correct translation of the
`source_text`
```

User Prompt

```
`source_text`: {source}
`target_text`: {target}
```

The model is trained on 60,346 training samples and achieves an overall accuracy of 82% on the validation set of 25,863 samples. A complete evaluation report is shown below in Table 9.

Label	Precision	Recall	F1	Support
Bad	70	73	71	2066
Partially Correct	76	63	69	7704
Entirely Correct	87	93	90	16093

Table 9: Translation Quality Classifier performance metrics

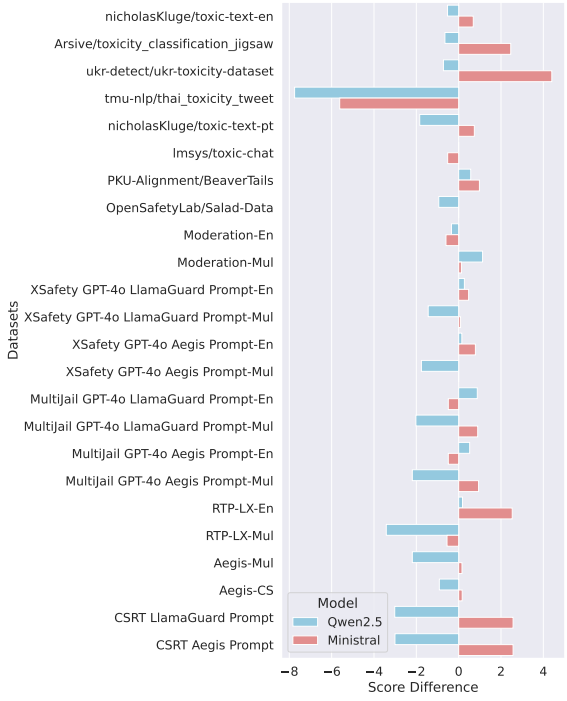

Figure 9: Performance difference on removing low-quality data. **Takeaway**: *Removal of low-quality training data does not necessarily improve model performance.*

Removal of low-quality training data does not necessarily improve model performance. Intuitively, the presence of poor-quality translated data should harm model performance. However, PG models show contrastive trends when low-quality samples are removed from the training data mix (Figure 9). The performance of `Qwen2.5` degrades for most datasets, whereas the performance of `Ministral` improves. The performance degradation in the case of `Qwen2.5` can be attributed to noisy samples in safety and toxicity evaluation datasets. Harmful text is considered to belong to low-quality data; web-crawls implement word blocklist filters to enhance data quality (Dodge et al., 2021). Thus, we hypothesize that the noise induced by poor translations bridges the gap between training and evaluation data, thus leading to performance improvement.