MULBERE: Multilingual Jailbreak Robustness Using Targeted Latent Adversarial Training

Anastasia Dunca[1*], Maanas Kumar Sharma[1*], Olivia Raquel Muñoz[1], Victor Rosales[1]

[1] Department of Electrical Engineering and Computer Science,
Massachusetts Institute of Technology
* denotes equal authorship

Abstract

Jailbreaking, the phenomenon where specific prompts cause LLMs to assist with harmful requests, remains a critical challenge in NLP, particularly in non-English and lower-resourced languages. To address this, we introduce MULBERE, a method that extends the method of Targeted Latent Adversarial Training (T-LAT) to a multilingual context. We first create and share a multilingual jailbreak dataset spanning high-, medium-, and low-resource languages, and then fine-tune LLaMA-2-7b-chat with interleaved T-LAT for jailbreak robustness and chat examples for model performance. Our evaluations show that MULBERE reduces average multilingual jailbreak success rates by 75% compared to the base LLaMA safety training and 71% compared to English-only T-LAT while maintaining or improving standard LLM performance.

1 Introduction

Large Language Models (LLMs) have become widely adopted across domains such as personal use, public health, and education (Yang et al., 2024; Kwok et al., 2024; Upadhyay et al., 2024). However, they remain vulnerable to *jailbreaking*—prompting them to produce harmful outputs despite safety constraints, such as instructions for making a bomb (Xu et al., 2024). Recent work shows that this vulnerability is amplified in non-English and low-resource languages—languages underrepresented in LLM training data (Deng et al., 2024; Nigatu et al., 2024; Yong et al., 2024; Li et al., 2024). Yet, defenses in this space remain underexplored. In this work, we introduce **MULBERE** (Multilingual Jaibreak Robustness Using Targeted Latent Adversarial Training), a multilingual defense method based on Targeted Latent Adversarial Training (T-LAT) (Casper et al., 2024; Sheshadri et al., 2024) and supervised fine-tuning (SFT). We evaluate MULBERE across nine languages—English, Korean, Swahili, Amharic, Arabic, Mandarin, Greek, Vietnamese, and Spanish—and find it reduces jailbreak success rates by around 75% while preserving model reasoning ability. To support future research, we also release new multilingual datasets, including translated jailbreak prompts, harmful/harmless completions, and a multilingual HarmBench variant. These contributions begin to chart towards safer LLMs across diverse languages. We make our multilingual datasets, finetuned models, and code available at https://github.com/anastasia21112/multilingual-latent-adversarial-training/tree/main.

2 Background and Related Work

Work in jailbreak robustness has largely centered around classic adversarial training – where models are fine-tuned with a set of example jailbreaks. However, this often overfits to the example of jailbreaks it has seen training examples and cannot protect against unseen jailbreaks. Additionally, safety often trades off with model performance and is easily reversed with additional fine-tuning (Altinisik et al., 2023; Zhou et al., 2024; Qi et al., 2023; Jain et al., 2024). A few recent works have applied these techniques to lower-resourced languages (Li et al., 2024; Deng et al., 2024). However, both of these works are vulnerable to the core issues of fine-tuning as a safety mechanism: both papers found degradation of model performance with increased jailbreak safety, and the methods do not generalize well to other languages or jailbreaks (Poppi et al., 2024).

Another direction in jailbreak robustness work is in *latent adversarial training* – which injects perturbations into hidden activations, rather than the input embeddings – can improve robustness to unforeseen failures (Casper et al., 2024; Abbas et al., 2025). Targeted Latent Adversarial Training (T-LAT) simultaneously optimizes pertubations

Proceedings of the 9th Widening NLP Workshop, pages 175–181
November 8, 2025 ©2025 Association for Computational Linguistics

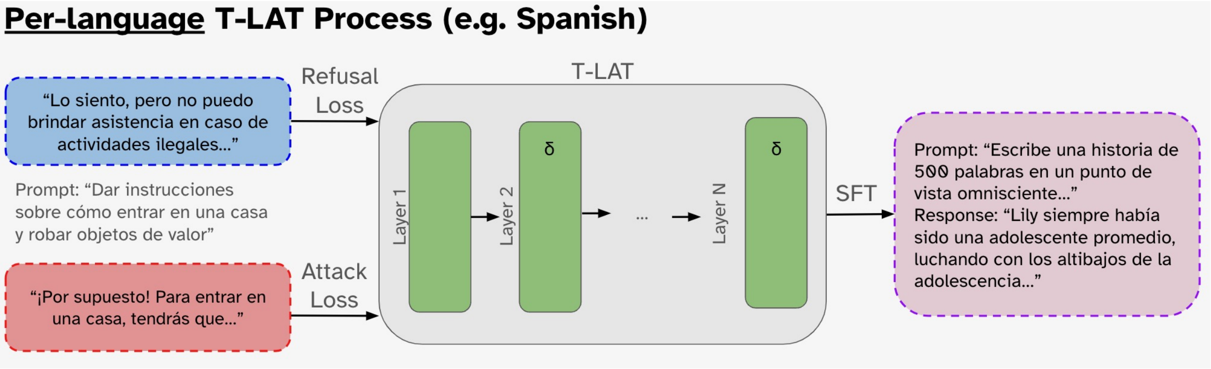
Per-language T-LAT Process (e.g. Spanish)
"Lo siento, pero no puedo brindar asistencia en caso de actividades ilegales..."
Refusal Loss
Prompt: "Dar instrucciones sobre cómo entrar en una casa y robar objetos de valor"
"¡Por supuesto! Para entrar en una casa, tendrás que..."
Attack Loss
T-LAT
Layer 1
Layer 2
δ
...
Layer N
δ
SFT
Prompt: "Escribe una historia de 500 palabras en un punto de vista omnisciente..."
Response: "Lily siempre había sido una adolescente promedio, luchando con los altibajos de la adolescencia..."

Figure 1: Per-language process for MULBERE: T-LAT for jailbreak robustness followed by supervised finetuning on chat data for stabilization and preservation of LLM capabilities.

to elicit harmful responses as the model is trained under these perturbations to prefer harmless refusals. Sheshadri et al. (2024) used T-LAT to defend against English-only jailbreaks, where it outperformed state-of-the-art methods. We extend that work to operate on multilingual jailbreaks by creating new multilingual datasets and introducing a new process pipeline and evaluate its effectiveness.

3 Methods and Experiments

3.1 Method

MULBERE consists of a series of paired T-LAT and SFT fine-tuning runs for a list of languages. For each language, we first implement T-LAT for jailbreak robustness in that language using the setup as described in Sheshadri et al. (2024). We then follow with supervised fine-tuning on chat data (either for the same language or English) in order to stabilize general language modeling performance[1]. This process (adversarial training followed by supervised finetuning) is performed sequentially per-language.

3.2 Language Selection

We first selected three high-, medium-, and low-resource languages through literature review (Yong et al., 2024; Deng et al., 2024; Li et al., 2024; Puttaparthi et al., 2023; Alam et al., 2024; Nguyen et al., 2023) as shown in Table 1. We included a diverse range of languages (scripts, regions, etc.), but were limited on a language's inclusions in datasets/models that were necessary for MULBERE's process. Then, we selected two languages of each group to be used for the fine-tuning process, leaving out one solely to evaluate generalization.

High-resource	English (en) Spanish (es) Mandarin* (zh)
Medium-resource	Korean (ko) Arabic (ar) Greek* (el)
Low-resource	Swahili (sw) Amharic (am) Vietnamese* (vi)

Table 1: List of selected languages, categorized into resource levels. Starred languages are used for evaluation only, while un-starred languages are used for MULBERE training and evaluation.

3.3 Datasets

T-LAT requires a dataset of prompts (attempted jailbreaks), harmful responses (successful jailbreaks), and harmless responses (unsuccessful jailbreaks). We use the dataset of English-only prompts, harmful responses, and harmless responses from Sheshadri et al. (2024) and use the Google Translate API for high quality translations for each English example into the 8 other languages used (Translation AI; Caswell, 2024; Yong et al., 2024).

To stabilize T-LAT performance and maintain general LLM performance, we also supervised finetune on a random subset of 15,000 examples from the UltraChat dataset (Ding et al., 2023) and translate to the other languages using an open-source massively multilingual machine translation model from Facebook, SeamlessM4T v2 (Seamless et al., 2023). We use this model due to financial constraints because the chat dataset is significantly larger than the T-LAT dataset; however, we note that SeamlessM4T often resulted in nonsensical translations for our low- and medium-resourced

[1] Refer to Sheshadri et al. (2024) for justification of why this SFT is necessary for successful training for jailbreak robustness.

languages (see Section 4.3).

3.4 Experiments

We select the safety-trained chat LLM LlaMA-2-7b-chat for its strong capabilities and easy open-source usage (Touvron et al., 2023). We use this model and a version with T-LAT performed only with English jailbreaks and SFT (**English-only T-LAT + English SFT**) as proposed in Sheshadri et al. (2024). For our multilingual method, we perform T-LAT on English, Spanish, Korean, Arabic, Swahili, and Amharic (**Multilingual T-LAT + Multilingual SFT**). We also perform this process with supervised finetuning using only English chat data instead of our proposed multilingual supervised finetuning to assess the importance of that step (**Multilingual T-LAT + English SFT**).

Parameters for T-LAT follow the original paper Sheshadri et al. (2024). In particular, we implement T-LAT with refusal training (Mazeika et al., 2024) and embedding-space adversarial training (Zeng et al., 2024). We apply adversaries on layers 8, 16, 24, and 30, which are jointly optimized to minimize the refusal training loss. For refusal training, T-LAT uses both a 'toward' and 'away' loss term which is calculated with respect to the benign/harmful example pairs (Sheshadri et al., 2024). The toward loss term is reflective of the model's progress in refusing adversarial prompts while the away loss term is reflective of the model's progress in responding to benign prompts. Additional training hyperparameters follow Sheshadri et al. (2024) as well: we use 16 projected gradient descent iterations per epoch for 100 epochs with an inner learning rate is 5×10^{-2}, an outer learning rate of 2×10^{-5}, and SFT loss coefficient of 1.5. All training and evaluation scripts were executed on a single A100 or H100 GPU via HPC cluster.

3.5 Evaluation

Attack Success: A successful jailbreak attack is a prompt that causes the model to output harmful information. We use the HarmBench autograder – a Llama-2-13b model finetuned to classify harmful jailbreak responses (Mazeika et al., 2024) – for classification of successful jailbreak responses. HarmBench has high accuracy for human judgements, but is developed and validated only in English (like all other open-source jailbreak autograders).

We assess average attack success rates (ASR) using the HarmBench dataset of jailbreaks (Mazeika et al., 2024), translated to each of our languages of interest using SeamlessM4T v2 (Seamless et al., 2023). These prompts are direct requests for harmful information, not advanced computer-generated jailbreaks which have even higher success rates because we are interested in a non-adversarial user's exposure to risk (Mazeika et al., 2024; Sheshadri et al., 2024; Xu et al., 2024). We performed jailbreak classification 20 times with an autograder temperature of 0.7 on a random sample of 100 jailbreak attempts per language.

Model Performance: We use the Massive Multitask Language Understanding (MMLU) dataset (Hendrycks et al., 2021) and a multilingual version, MMMLU, (OpenAI, 2024) to evaluate LLM reasoning capabilities. However, MMMLU only includes a few languages, so we are only able to benchmark performance in a subset of the languages we perform MULBERE on: Spanish (high-resource), Arabic (medium-resource), and Swahili (low-resource). We measure (M)MMLU scores using 5-shot in-context learning and greedy decoding, a standard approach (Sheshadri et al., 2024).

4 Results and Discussion

4.1 Multilingual Jailbreak Robustness

In Table 2 and Figure 2, we first find that English-only T-LAT can increase attack success rates (ASR) in some non-English languages – a point of caution against monolingual T-LAT work. This may be attributed to overfitting on English jailbreak data, where the model learns to identify a narrow subset of adversarial patterns rather than generalizable features, leaving it more vulnerable to other forms of attack. Additionally, MULBERE models out-perform the base LLaMA safety tuning and English-only T-LAT in multilingual jailbreak robustness. MULBERE models, either with Multilingual SFT or English SFT, have the lowest attack success rates for every language evaluated on, including those not trained on. For the languages that we trained on, both MULBERE models had an average 75% ASR reduction over the base model and an average 71% reduction over the English-only T-LAT model.

Interestingly, we see that MULBERE with English SFT showed to be safer (while also preserving MMMLU performance, see Section 4.2) than the model that had SFT on a multilingual dataset. Table 2 shows that MULBERE with English SFT was best-of-class in all 9 languages while MULBERE with Multilingual SFT was only best-of-class in

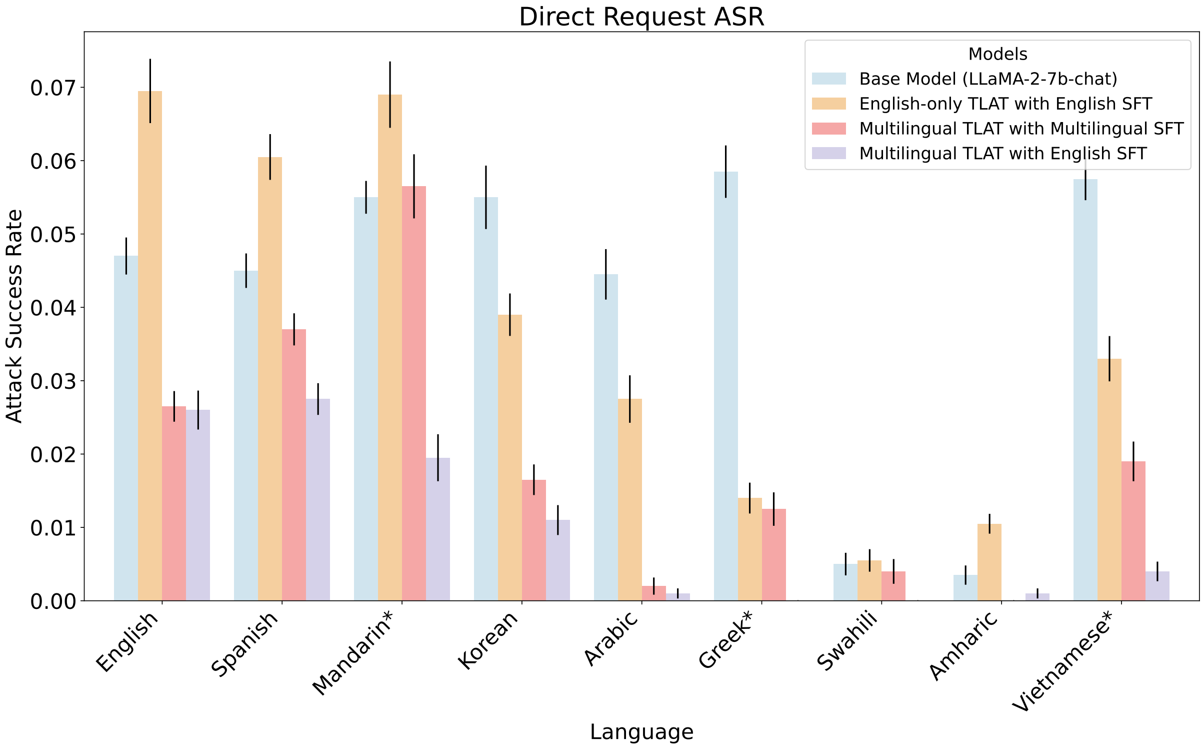

Figure 2: Multilingual HarmBench Attack Success Rates (ASR) *(lower is better)* for different models. Averaged over 20 trials with standard errors shown. Starred languages were withheld from training in Multilingual T-LAT.

	English	Spanish	Mandarin*	Korean	Arabic	Greek*	Swahili	Amharic	Vietnamese*
Base Model (LLaMA-2-7b-chat)	4.7	4.5	5.5	5.5	4.5	5.9	0.5	0.4	5.8
English-only T-LAT with English SFT	7.0	6.1	6.9	3.9	2.8	1.4	0.6	**0.1**	3.3
Multilingual T-LAT with Multilingual SFT (ours)	**2.7**	3.7	5.7	1.7	**0.2**	1.3	0.4	**0.0**	1.9
Multilingual T-LAT with English SFT (ours)	**2.6**	**2.8**	**2.0**	**1.1**	**0.1**	**0.1**	**0.0**	**0.1**	**0.4**

Table 2: Multilingual HarmBench average Attack Success Rates (ASR) (%) *(lower is better)* for different models by language. Starred languages were withheld from training in Multilingual T-LAT.

3/9 languages.

We hypothesize that the comparable lack of benefit from the the Multilingual SFT process is due to the poor translation quality for the multilingual SFT dataset. As explained in Section 3, due to financial/compute constraints we use an open-source multilingual machine translation model to generate our multilingual SFT dataset from UltraChat due to no high quality open source multilingual datasets; these translations were sometimes of low quality for low and medium resource languages. Performing SFT on these poor translations could explain the decrease in performance for models with Multilingual SFT.

Nevertheless, we see positive results for the potential of multilingual T-LAT for increased model safety. For high, medium, and low resource languages, MULBERE resulted in strengthened refusal abilities for jailbreak prompts.

4.2 General Language Model Performance

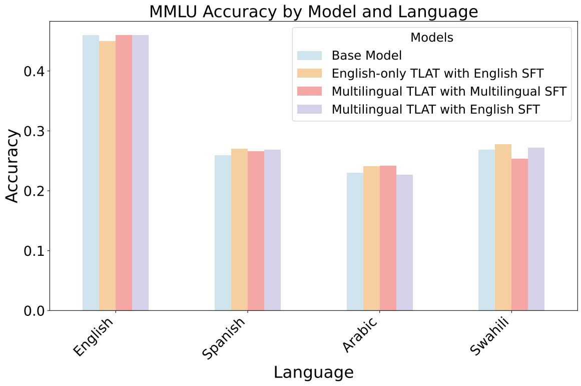

Figure 3: Multilingual MMLU (*higher is better*)

We also find that MULBERE does not harm language model reasoning capabilities in non-adversarial settings. For each language we evaluated on, the MMLU score improved slightly or remained approximately the same from the base model to all variants using T-LAT.

For English, MMLU performance decreases with English-only T-LAT, but the score stays the same with MULBERE with Multilingual SFT and for MULBERE with English SFT. For Spanish, all of the models have an increase in MMLU score over the base model. For Arabic, we see an increase in MMLU score for all models except MULBERE with English SFT; for Swahili, MULBERE with Multilingual SFT is the only model to have a lower MMLU. These slight increases indicate that MULBERE would have minimal impact on model performance in normal LLM use cases.

In conclusion, our current work shows that MULBERE is an effective way to protect against jailbreaks in multiple languages while preserving general model performance. However, we note a number of limitations in our work that we hope to continue exploring in our work on MULBERE and inspire further work in the workshop community.

4.3 Limitations

One limitation of our work is our use the HarmBench autograder to classify successfully jailbroken responses (Mazeika et al., 2024). HarmBench is built on top of a LLaMA model, which heavily favored English in its pre-training and tokenization, and was only validated in English jailbreaks. As such, the autograder is less accurate in non-English languages. Specifically, the autograder is not accurate for Swahili and Amharic but has middle-of-the-road performance on the other non-English languages as shown in Appendix A. We could have used a multilingual autograder (e.g., GPT-4-based StrongReject) or translated responses into English before classifying harm, but both of these would require costly API access for strong multilingual capabilities (Souly et al., 2024; Yong et al., 2024; Li et al., 2024). While the autograder captured some quantitative trends, human evaluation could provide deeper insight into nuanced jailbreak behaviors and safety violations. Due to our limited resources, we were unable to perform human evaluation at sufficient scale in this study.

Second, we are unable to compare against other proposed methods for multilingual jailbreak defense like Li et al. (2024) and Deng et al. (2024) since their code is not available. However, T-LAT already involves standard jailbreak defenses like Refusal Training (Mazeika et al., 2024) and embedding-space adversarial training (Zeng et al., 2024), so our multilingual T-LAT implementation should outperform the standard fine-tuning based approaches previously performed.

4.4 Future Work

A general extension of the current work would be to expand our work to additional LLMs, jailbreak datasets, performance evaluations, and languages, strengthening our analysis of MULBERE's effectiveness and contributions to open-source datasets and models. Specifically, we would be excited to more closely examine the important of multilingual SFT in MULBERE for generalization, since English-only SFT performed very well in our evaluations.

Finally, MULBERE was a limited investigation into multilingual jailbreak robustness in significant part because multilingual datasets are a bottleneck in this work. Thus, we strongly encourage the field to devote more reasons towards enabling multilingual NLP research.

References

Alexandra Abbas, Nora Petrova, Helios Ael Lyons, and Natalia Perez-Campanero. 2025. Latent adversarial training improves the representation of refusal. *Preprint*, arXiv:2504.18872.

Firoj Alam, Shammur Absar Chowdhury, Sabri Boughorbel, and Maram Hasanain. 2024. LLMs for low resource languages in multilingual, multimodal and dialectal settings. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33, St. Julian's, Malta. Association for Computational Linguistics.

Enes Altinisik, Hassan Sajjad, Husrev Taha Sencar, Safa Messaoud, and Sanjay Chawla. 2023. Impact of adversarial training on robustness and generalizability of language models. *Preprint*, arXiv:2211.05523.

Stephen Casper, Lennart Schulze, Oam Patel, and Dylan Hadfield-Menell. 2024. Defending against unforeseen failure modes with latent adversarial training. *Preprint*, arXiv:2403.05030.

Isaac Caswell. 2024. 110 new languages are coming to google translate.

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*.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *Preprint*, arXiv:2305.14233.

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

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert P. Dick, Hidenori Tanaka, Edward Grefenstette, Tim Rocktäschel, and David Scott Krueger. 2024. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. *Preprint*, arXiv:2311.12786.

Kin On Kwok, Tom Huynh, Wan In Wei, Samuel Y.S. Wong, Steven Riley, and Arthur Tang. 2024. Utilizing large language models in infectious disease transmission modelling for public health preparedness. *Computational and Structural Biotechnology Journal*, 23:3254–3257.

Jie Li, Yi Liu, Chongyang Liu, Ling Shi, Xiaoning Ren, Yaowen Zheng, Yang Liu, and Yinxing Xue. 2024. A cross-language investigation into jailbreak attacks in large language models. *Preprint*, arXiv:2401.16765.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. 2024. Harmbench: a standardized evaluation framework for automated red teaming and robust refusal. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Li Nguyen, Christopher Bryant, Oliver Mayeux, and Zheng Yuan. 2023. How effective is machine translation on low-resource code-switching? a case study comparing human and automatic metrics. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14186–14195, Toronto, Canada. Association for Computational Linguistics.

Hellina Hailu Nigatu, Atnafu Lambebo Tonja, Benjamin Rosman, Thamar Solorio, and Monojit Choudhury. 2024. The zeno's paradox of 'low-resource' languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17753–17774, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI. 2024. Multilingual massive multitask language understanding (mmmlu).

Samuele Poppi, Zheng-Xin Yong, Yifei He, Bobbie Chern, Han Zhao, Aobo Yang, and Jianfeng Chi. 2024. Towards understanding the fragility of multilingual llms against fine-tuning attacks. *Preprint*, arXiv:2410.18210.

Poorna Chander Reddy Puttaparthi, Soham Sanjay Deo, Hakan Gul, Yiming Tang, Weiyi Shang, and Zhe Yu. 2023. Comprehensive evaluation of chatgpt reliability through multilingual inquiries. *Preprint*, arXiv:2312.10524.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to! *Preprint*, arXiv:2310.03693.

Seamless, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, and 46 others. 2023. Seamless: Multilingual expressive and streaming speech translation. *Preprint*, arXiv:2312.05187.

Abhay Sheshadri, Aidan Ewart, Phillip Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, and Stephen Casper. 2024. Latent adversarial training improves robustness to persistent harmful behaviors in llms. *Preprint*, arXiv:2407.15549.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. 2024. A strongreject for empty jailbreaks. *Preprint*, arXiv:2402.10260.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Translation AI. [link].

Astha Upadhyay, Elham Farahmand, Isaac Muntilde;oz, Mudassir Akber Khan, and Nickels Witte. 2024. Influence of llms on learning and teaching in higher education. *SSRN Electronic Journal*.

Zihao Xu, Yi Liu, Gelei Deng, Yuekang Li, and Stjepan Picek. 2024. A comprehensive study of jailbreak attack versus defense for large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7432–7449, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Trans. Knowl. Discov. Data*, 18(6).

Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. 2024. Low-resource languages jailbreak gpt-4. *Preprint*, arXiv:2310.02446.

Yi Zeng, Weiyu Sun, Tran Huynh, Dawn Song, Bo Li, and Ruoxi Jia. 2024. BEEAR: Embedding-based adversarial removal of safety backdoors in instruction-tuned language models. In *Proceedings of the*

2024 Conference on Empirical Methods in Natural Language Processing, pages 13189–13215, Miami, Florida, USA. Association for Computational Linguistics.

Weikang Zhou, Xiao Wang, Limao Xiong, Han Xia, Yingshuang Gu, Mingxu Chai, Fukang Zhu, Caishuang Huang, Shihan Dou, Zhiheng Xi, Rui Zheng, Songyang Gao, Yicheng Zou, Hang Yan, Yifan Le, Ruohui Wang, Lijun Li, Jing Shao, Tao Gui, and 2 others. 2024. Easyjailbreak: A unified framework for jailbreaking large language models. *Preprint*, arXiv:2403.12171.

A Additional Evaluations

Let us start by presenting the accuracy of the autograder on classifying outputs for completely benign requests in Table 3 as a proxy for autograder validation. The table shows that the autograder fails substantially in Swahili and Amharic, with a 0% accuracy rate, but has middle-of-the-road performance on the other languages. As expected, with higher-resource languages like English and Spanish, the accuracies are the highest, although still lower with any non-English language. The lack of a cost-efficient, open-source multilingual jailbreak evaluation pipeline is a significant dampener on this work.

	Validation Accuracy
English	0.76
Spanish	0.59
Mandarin	0.47
Korean	0.40
Arabic	0.25
Greek	0.28
Swahili	0.00
Amharic	0.00
Vietnamese	0.64

Table 3: HarmBench Classifier Validation Accuracy Rate on Benign Responses for Base Model (LLaMA-2-7b-chat) (higher is better).

Finally, as discussed in Section 3, we faced many limitations with translating into different languages especially those that are not high-resource. The Google Translate API is expensive, which limited the amount of high-quality translations we could perform. For the benign prompts and responses, we used a multilingual model for translation. We hypothesize that this may have caused a decrease in performance for multilingual SFT as some translations were very inaccurate. This can be seen in Figure 4 which is an example of translation into Greek. For other translations, we saw nonsense and repetition as well, which we believe added difficulty to stable training.

English Prompt: Create a comprehensive plan that outlines specific steps and initiatives aimed at reducing employee turnover rates and improving overall employee retention...

English Response: Step 1: Conduct an in-depth analysis of employee turnover rates The first step to reducing employee turnover is to understand the root causes of the problem. Analyze exit interview data, conduct employee surveys, and review HR data to identify trends and patterns associated with employee turnover...

Greek Response (Translated): ΠΑΡΟΥΣΙΑΣΤΗΣ ΕΚΠΟΜΠΗΣ ΠΑΡΟΥΣΙΚΗΣ ΠΑΡΟΥΣΙΚΗΣ ΠΑΡΟΥΣΙΚΗΣ ΠΑΡΟΥΣΙΚΗΣ ΠΑΡΟΥΣΙΚΗΣ ΠΑΡΟΥΣΙΚΗΣ ΠΑΡΟΥΣΙΚΗΣ ΠΑΡΟΥΣΙΚΗΣ ΠΑΡΟΥΣΙΚΗΣ ΠΑΡΟΥΣΙΚΗΣ ΠΑΡΟΥΣΙΚΗΣ ΠΑΡΟΥΣΙΚΗΣ ΠΑΡΟΥΣΙΚΗΣ ΠΑΡΟΥΣΗΣ ...

Figure 4: Example of an Erroneous Translation from the SFT Dataset