GuardBench: A Large-Scale Benchmark for Guardrail Models

Elias Bassani
European Commission
Joint Research Centre
Ispra, Italy
elias.bassani@ec.europa.eu

Ignacio Sanchez
European Commission
Joint Research Centre
Ispra, Italy
ignacio.sanchez@ec.europa.eu

Abstract

Generative AI systems powered by Large Language Models have become increasingly popular in recent years. Lately, due to the risk of providing users with unsafe information, the adoption of those systems in safety-critical domains has raised significant concerns. To respond to this situation, input-output filters, commonly called guardrail models, have been proposed to complement other measures, such as model alignment. Unfortunately, the lack of a standard benchmark for guardrail models poses significant evaluation issues and makes it hard to compare results across scientific publications. To fill this gap, we introduce GuardBench, a large-scale benchmark for guardrail models comprising 40 safety evaluation datasets. To facilitate the adoption of GuardBench, we release a Python library providing an automated evaluation pipeline built on top of it. With our benchmark, we also share the first large-scale prompt moderation datasets in German, French, Italian, and Spanish. To assess the current state-of-the-art, we conduct an extensive comparison of recent guardrail models and show that a general-purpose instruction-following model of comparable size achieves competitive results without the need for specific fine-tuning.[1]

1 Introduction

In the recent years, Generative AI systems have become increasingly popular thanks to the advanced capabilities of Large Language Models (LLMs) (OpenAI, 2023). Those systems are in the process of being deployed in a range of high-risk and safety-critical domains such as healthcare (Meskó and Topol, 2023; Zhang and Boulos, 2023), education (Baidoo-Anu and Ansah, 2023; Qadir, 2023), and finance (Chen et al., 2023). As AI systems advance and are more extensively integrated into various application domain, it is crucial to ensure that their usage is secure, responsible, and compliant with the applicable AI safety regulatory framework.

Particular attention has been paid to chatbot systems based on LLMs, as they can potentially engage in unsafe conversations or provide users with information that may harm their well-being. Despite significant efforts in aligning LLMs to human values (Wang et al., 2023b), users can still misuse them to produce hate speech, spam, and harmful content, including racist, sexist, and other damaging associations that might be present in their training data (Wei et al., 2023). To alleviate this situation, explicit safeguards, such as input-output filters, are becoming fundamental requirements for safely deploying systems based on LLMs, complementing other measures such as model alignment.

Very recently, researchers have proposed the adoption of the so-called guardrail models to moderate user prompts and LLM-generated responses (Inan et al., 2023; Ghosh et al., 2024; Li et al., 2024). Given the importance of those models, their evaluation plays a crucial role in the Generative AI landscape. Despite the availability of a few datasets for assessing guardrail models capabilities, such as the OpenAI Moderation Dataset (Markov et al., 2023) and BeaverTails (Ji et al., 2023), we think there is still need for a large-scale benchmark that allows for a more systematic evaluation.

We aim to fill this gap by providing the scientific community with a large-scale benchmark comprising several datasets for prompts and responses safety classification. To facilitate the adoption of our proposal, we release a Python library that provides an automated evaluation pipeline built on top of the benchmark itself. Moreover, we share the first large-scale multi-lingual prompt moderation datasets, thus overcoming English-only evaluation. Finally, we conduct the first extensive comparison of recent guardrail models, aiming at shedding some light on the state-of-the-art and show a general-purpose instruction-following model of

[1] https://github.com/AmenRa/guardbench

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 18393–18409
November 12-16, 2024 ©2024 Association for Computational Linguistics

comparable size achieves competitive results without the need for specific fine-tuning.

Our contributions can be summarized as follows:

- We introduce a large-scale benchmark for guardrail models evaluation composed of 40 datasets, overcoming models comparison limited to a few datasets.
- We share the first prompt safety datasets in German, French, Italian, and Spanish, comprising more than 31k prompts each.
- We share a novel AI response evaluation dataset comprising 22k question-answer pairs.
- We release a Python library to facilitate the adoption of the proposed benchmark.
- We conduct the first extensive evaluation of guardrail models, comparing 13 models on 40 prompts and conversations safety datasets.

2 Related Work

In this section, we discuss previous work related to our benchmark. Firstly, we discuss the moderation of user-generated content. Secondly, we introduce the moderation of human-AI conversations.

2.1 Moderation of User-Generated Content.

The most related task to the one of our benchmark is the moderation of user-generated content, which has received significant attention in the past decade. Many datasets for the evaluation of moderation models have been proposed by gathering user-generated content from social networks and online forums, such as Twitter, Reddit, and others (Basile et al., 2019; Kennedy et al., 2022; Davidson et al., 2017; ElSherief et al., 2021; Kennedy et al., 2020; Zampieri et al., 2019; Guest et al., 2021; Grimminger and Klinger, 2021; Sap et al., 2020; de Gibert et al., 2018). However, the task of moderating human-AI conversations is different in nature to that of moderating user-generated content. First, the texts produced in human-AI conversations differ from that generated by users on online social platforms. Second, LLM-generated content further differs from that generated by users in style and length (Herbold et al., 2023; Gao et al., 2023). Finally, the type of unsafe content in content moderation datasets is typically limited to hate and discrimination, while the unsafe content potentially present in human-AI conversation is much broader, ranging from weapons usage to cybersecurity attacks and self-harm (Inan et al., 2023).

2.2 Moderation of Human-AI Conversations.

The moderation of human-AI conversations comprises both the moderation of human-generated and LLM-generated content. In this context, users ask questions and give instructions to LLMs, which answer the user input. Unfortunately, LLMs may engage in offensive conversations (Lee et al., 2019; Curry and Rieser, 2018) or generate unsafe content in response to the user requests (Dinan et al., 2019). To moderate such conversations, guardrail models have recently been proposed (Inan et al., 2023; Ghosh et al., 2024; Li et al., 2024), aiming to enforce safety in conversational AI systems or evaluate it before deployment (Vidgen et al., 2024; Li et al., 2024). Our work focus on both the moderation of user prompts and LLM responses. Specifically, we collect and extend several datasets related to LLM safety, providing the scientific community with a large-scale benchmark for the evaluation of guardrail models.

3 Benchmark Composition

In this section, we introduce the benchmark we have built by collecting several datasets from previous works and extending them through data augmentation. To decide which datasets to include in our evaluation benchmark, we first conducted a literature review and consulted SafetyPrompts[2] (Röttger et al., 2024). We considered over 100 datasets related to LLM safety. To narrow down the initial list of datasets and identify those best suited for our evaluation purposes, we defined inclusion and exclusion criteria, which we present in Section 3.1. As many of these datasets were not proposed to evaluate guardrail models, we repurposed them to our needs as they already contained safety information. We include 35 datasets from previous works in our benchmark, which can be broadly categorized as prompts (instructions, question, and statements) or conversations (single-turn and multi-turn), where the object to be moderated is the final utterance. Due to the lack of non-English datasets (Röttger et al., 2024), we augmented those available through automatic translation, providing the scientific community with the first prompts safety evaluation sets for guardrail models in German, French, Italian, and Spanish. We detail such process in Section 3.3. Finally, as described in Section 3.4, we generate safe and unsafe responses to unsafe questions and instructions

[2] `https://safetyprompts.com`

from previous works to obtain a novel large-scale conversational dataset for our evaluation. The final list of datasets comprised in our benchmark is presented in Table 1.

3.1 Inclusion and Exclusion Criteria

In this section, we introduce inclusion and exclusion criteria adopted for selecting safety datasets.

- We include datasets comprising text chat between users and AI assistants, open-ended questions and instructions, and other texts that can be expressed in a prompt format.
- We include datasets with safety labels that resembles or fall within generally acknowledged harm categories (Vidgen et al., 2024).
- We include public datasets available on GitHub[3] and HuggingFace's Datasets (Lhoest et al., 2021).
- We include datasets with permissive licenses, such as MIT, CC BY(-NC), and Apache 2.0.
- Due to the lack of non-English datasets (Röttger et al., 2024), we initially consider only datasets in English.
- We exclude content moderation datasets from social networks and online forums. As explained in Section 2.1, their content differ from both user prompts and LLM responses.
- We exclude safety evaluation datasets that cannot be straightforwardly repurposed for the evaluation of guardrail models, such as multi-choice datasets (Zhang et al., 2023) and completion datasets (Gehman et al., 2020).
- We exclude datasets whose samples' safety labels were computed by automated tools (e.g, Perspective API[4], OpenAI Moderation API[5]), such as RealToxicityPrompts (Gehman et al., 2020), LMSYS-Chat-1M (Zheng et al., 2023), and the toxicity dataset comprised in DecodingTrust (Wang et al., 2023a).
- We exclude datasets that need to be built from scratch, such as AdvPromptSet, (Esiobu et al., 2023) or protected by password, such as FairPrism (Fleisig et al., 2023).
- We exclude datasets for jail-breaking and adversarial robustness evaluation, as jail-breaking and adversarial attacks are not the main focus of our work. However, we do include the unsafe prompts contained in those datasets (without jail-breaking or adversarial texts) as they are relevant to out work.

3.2 Classification Task

For our benchmark, we consider the safe/unsafe binary classification task for the following reasons. Firstly, due to the lack of a generally accepted taxonomy of unsafe content (Vidgen et al., 2024) and differences in the labeling procedures of previous works, we are unable to map the unsafe content categories of every dataset to a reference taxonomy. Secondly, several datasets lack this information and only provide implicit safety categorization of the shared samples, i.e., they are all unsafe by construction. Therefore, we binarize the labels of the available datasets into safe/unsafe. By inspecting previous works' categories of harm, we ensure that all the datasets' unsafe samples fall within generally acknowledged harm categories, such as hate, discrimination, violence, weapons, adult content, child exploitation, suicide, self-harm, and others. Despite specific labeling differences, we find all the selected datasets to adhere to a shared safe/unsafe distinction, corroborating our design choice. Appendix A.1 details the label conversion process for each of the chosen datasets.

3.3 Multilingual Augmentation

As reported by Röttger et al. (2024), there is a lack non-English datasets for LLM safety evaluation. To overcome this limitation and conduct preliminary experiments with guardrail models on non-English texts, we translate the datasets of prompts in our benchmark to several languages. Specifically, by relying on Google's MADLAD-400-3B-MT (Kudugunta et al., 2023), we translate 31k prompts into German, French, Italian, and Spanish. To ensure the quality of the translations, we asked native speakers to evaluate four prompts from each translated dataset ($\sim$ 100 prompts per language) and score them on a five-point Likert scale (Likert, 1932) where one means that the translation is wrong and five means that the translation is perfect. Our annotators judged that the average translation quality exceed four points. We add the obtained datasets to GuardBench as PromptsDE, PromptsFR, PromptsIT, and PromptsES. The list of datasets used to derive our multi-lingual datasets is available in Appendix A.2.

[3] https://github.com

[4] https://www.perspectiveapi.com

[5] https://platform.openai.com/docs/guides/moderation

Dataset	Category	Sub-category	Total	Unsafe	Labels	Source	Purpose	License	Reference
AdvBench Behaviors	Prompts	Instructions	520	100%	Auto	LLM	General Safety	MIT	Zou et al. (2023)
HarmBench Behaviors	Prompts	Instructions	320	100%	Auto	Human	General Safety	MIT	Mazeika et al. (2024)
I-CoNa	Prompts	Instructions	178	100%	Manual	Human	Hate	CC BY-NC 4.0	Bianchi et al. (2023)
I-Controversial	Prompts	Instructions	40	100%	Manual	Human	Controversial Topics	CC BY-NC 4.0	Bianchi et al. (2023)
I-MaliciousInstructions	Prompts	Instructions	100	100%	Auto	Mixed	General Safety	CC BY-NC 4.0	Bianchi et al. (2023)
I-Physical-Safety	Prompts	Instructions	200	50%	Manual	Human	Physical Safety	CC BY-NC 4.0	Bianchi et al. (2023)
MaliciousInstruct	Prompts	Instructions	100	100%	Auto	LLM	General Safety	MIT	Huang et al. (2023)
MITRE	Prompts	Instructions	977	100%	Manual	Mixed	Cybersecurity	MIT	Bhatt et al. (2024)
StrongREJECT Instructions	Prompts	Instructions	213	100%	Manual	Human	General Safety	MIT	Souly et al. (2024)
TDCRedTeaming Instructions	Prompts	Instructions	50	100%	Manual	Human	General Safety	MIT	Mazeika et al. (2023)
CatQA	Prompts	Questions	550	100%	Auto	LLM	General Safety	Apache 2.0	Bhardwaj et al. (2024)
Do Anything Now Questions	Prompts	Questions	390	100%	Auto	LLM	General Safety	MIT	Shen et al. (2023)
DoNotAnswer	Prompts	Questions	939	100%	Auto	LLM	General Safety	Apache 2.0	Wang et al. (2024)
HarmfulQ	Prompts	Questions	200	100%	Auto	LLM	General Safety	MIT	Shaikh et al. (2023)
HarmfulQA Questions	Prompts	Questions	1960	100%	Auto	LLM	General Safety	Apache 2.0	Bhardwaj and Poria (2023)
HEx-PHI	Prompts	Questions	330	100%	Manual	Human	General Safety	Custom	Qi et al. (2023)
XSTest	Prompts	Questions	450	44%	Manual	Human	Exaggerated Safety	CC BY 4.0	Röttger et al. (2023)
AdvBench Strings	Prompts	Statements	574	100%	Auto	LLM	General Safety	MIT	Zou et al. (2023)
DecodingTrust Stereotypes	Prompts	Statements	1152	100%	Manual	Template	Stereotypes	CC BY-SA 4.0	Wang et al. (2023a)
DynaHate	Prompts	Statements	4120	55%	Manual	Human	Hate	Apache 2.0	Vidgen et al. (2021)
HateCheck	Prompts	Statements	3728	69%	Manual	Template	Hate	CC BY 4.0	Röttger et al. (2021)
Hatemoji Check	Prompts	Statements	593	52%	Manual	Template	Hate w/ emojis	CC BY 4.0	Kirk et al. (2022)
SafeText	Prompts	Statements	1465	25%	Manual	Human	Physical Safety	MIT	Levy et al. (2022)
ToxiGen	Prompts	Statements	940	43%	Manual	LLM	Implicit Hate	MIT	Hartvigsen et al. (2022)
AART	Prompts	Mixed	3269	100%	Auto	LLM	General Safety	CC BY 4.0	Radharapu et al. (2023)
OpenAI Moderation Dataset	Prompts	Mixed	1680	31%	Manual	Human	General Safety	MIT	Markov et al. (2023)
SimpleSafetyTests	Prompts	Mixed	100	100%	Manual	Human	General Safety	CC BY 4.0	Vidgen et al. (2023)
Toxic Chat	Prompts	Mixed	5083	7%	Manual	Human	General Safety	CC BY-NC 4.0	Lin et al. (2023)
BeaverTails 330k	Conversations	Single-Turn	11088	55%	Manual	Mixed	General Safety	MIT	Ji et al. (2023)
Bot-Adversarial Dialogue	Conversations	Multi-Turn	2598	36%	Manual	Mixed	Hate	Apache 2.0	Xu et al. (2021)
ConvAbuse	Conversations	Multi-Turn	853	15%	Manual	Mixed	Hate	CC BY 4.0	Curry et al. (2021)
DICES 350	Conversations	Multi-Turn	350	50%	Manual	Mixed	General Safety	CC BY 4.0	Aroyo et al. (2023)
DICES 990	Conversations	Multi-Turn	990	16%	Manual	Mixed	General Safety	CC BY 4.0	Aroyo et al. (2023)
HarmfulQA	Conversations	Multi-Turn	16459	45%	Auto	LLM	General Safety	Apache 2.0	Bhardwaj and Poria (2023)
ProsocialDialog	Conversations	Multi-Turn	25029	60%	Manual	Mixed	General Safety	CC BY 4.0	Kim et al. (2022)
PromptsDE	Prompts	Mixed	30852	61%	Mixed	LLM	General Safety	Custom	Our
PromptsFR	Prompts	Mixed	30852	61%	Mixed	LLM	General Safety	Custom	Our
PromptsIT	Prompts	Mixed	30852	61%	Mixed	LLM	General Safety	Custom	Our
PromptsES	Prompts	Mixed	30852	61%	Mixed	LLM	General Safety	Custom	Our
UnsafeQA	Conversations	Single-Turn	22180	50%	Auto	Mixed	General Safety	Custom	Our

Table 1: List of benchmark datasets. Category and Sub-category indicate the primary and the specific text categories, respectively. Total and Unsafe report the number of samples in the test sets and the percentage of unsafe samples, respectively. Labels indicate whether labels where obtained by manual annotation (Manual) or by dataset construction (Auto). Source indicates whether a dataset is based on human-generated texts (Human), machine-generated texts (LLM), a mix of the two (Mixed), or was obtained through templating (Template). Purpose indicates the safety area addressed by the datasets. In this case, General Safety means the dataset covers multiple categories of harm, from hate, discrimination, and violence to cybersecurity and self-harm.

3.4 Answering Unsafe Prompts

Given the number of (unanswered) unsafe questions and instructions from previous works, we propose a novel single-turn conversational dataset built by generating responses with a publicly available uncensored model.[6] Specifically, by controlling the model's system prompt, we generate 22k safe and unsafe responses to the available unsafe questions and instructions. A system prompt is a way to provide context, instructions, and guidelines to the model before prompting it. Using a system prompt, we can set the role, personality, tone, and other relevant information that helps the model behave as expected, thus allowing us to control the generation of safe and unsafe responses. In the case of safe responses, we also inform the model that the requests to answer are from malicious users and instruct the model to provide helpful and pro-social responses (Kim et al., 2022). This way, we limit refusals and ensure the model does not provide unsafe information when we do not want it to do so. To ensure response quality, we manually checked a sample of the produced answers, finding that the employed model was surprisingly good at generating the expected answers. We add the obtained dataset to our benchmark under the name of UnsafeQA. The list of datasets used to derive UnsafeQA is available in Appendix A.2.

3.5 Software Library

GuardBench is accompanied by a Python library with the same name that we hope will facilitate the adoption of our benchmark as a standard for

[6] https://huggingface.co/cognitivecomputations/dolphin-2.9.1-yi-1.5-34b

guardrail models evaluation. The main design principles behind the implementation of our Python library are as follows: 1) reproducibility, 2) usability, 3) automation, and 4) extendability. As exemplified in Listing 1, the library provides a predefined evaluation pipeline that only requires the user to provide a moderation function. The library automatically downloads the requested datasets from the original repositories, converts them in a standardized format, moderates prompts and conversations with the moderation function provided by the user, and ultimately saves the moderation outcomes in the specified output directory for later inspections. This way, users can focus on their own moderation approaches without having to worry about the evaluation procedure. Moreover, by sharing models' weights and moderation functions, guardrail models evaluation can be easily reproduced across research labs, thus improving research transparency. To this extend, our Python library also offers the possibility of building comparison tables and export them in LaTeX, ready for scientific publications. Finally, the user can import new datasets to extend those available out-of-the-box. Further information and tutorials are available on GuardBench's official repository. We also release the code to reproduce the evaluation presented in Sections 4 and 5.

```
from guardbench import benchmark

benchmark(
    # Moderation function provided by the user.
    moderate,
    model_name="moderator",
    out_dir="results",
    batch_size=32,
    datasets="all",
)
```

Listing 1: GuardBench API.

4 Experimental Setup

In this section, we introduce the experimental setup adopted to answer the following research questions:

RQ1 What is the best model at moderating user prompts?

RQ2 What is the best model at moderating human-AI conversations?

RQ3 How does available models perform on languages other than English?

RQ4 How does content moderation policies affect models' effectiveness?

To answer the research questions **RQ1** and **RQ2** we compare the effectiveness of several models at classifying prompts and conversation utterances as safe or unsafe. Then, to answer **RQ3**, we compare the models on our newly introduced multi-lingual prompt datasets, described in Section 3.3. Finally, we evaluate the importance of moderation policies by comparing the results of a general-purpose LLM with different policies to answer **RQ4**.

In the following sections, we introduce the models we have compared (Section 4.1) and discuss the evaluation metrics chosen to assess the models' effectiveness (Section 4.2) before presenting the results in Section 5.

4.1 Models

In this section, we introduce the models that we evaluated against our large-scale benchmark. We consider several open-weight models, including recent guardrail models, content moderation models often employed in real-world applications, and instruction-tuned general-purpose LLM prompted for content moderation. We consider the latter to evaluate their out-of-the-box capabilities in detecting unsafe prompts and responses. The major differences between guardrail models and content moderation models are that the first are meant to moderate human-AI conversations while the latter were trained on content from online social platforms. Moreover, guardrail models are usually prompted by providing them a content moderation policy, i.e., a list of unsafe content categories, while available content moderation models do not take advantage of such mechanism. The list of all the considered models is presented below. Further information are provided in Table 2.

- **Llama Guard:** guardrail model based on LLama 2 7B (Touvron et al., 2023) proposed by Inan et al. (2023).
- **Llama Guard 2:** updated version of Llama Guard based on LLama 3 8B[7].
- **Llama Guard Defensive:** Llama Guard additionally fine-tuned by Ghosh et al. (2024) with a strict content moderation policy.
- **Llama Guard Permissive:** Llama Guard additionally fine-tuned by Ghosh et al. (2024) with a permissive content moderation policy.
- **MD-Judge:** guardrail model obtained by fine-tuning Mistral 7B (Jiang et al., 2023)

[7] https://ai.meta.com/blog/meta-llama-3

on BeaverTails330K (Ji et al., 2023), Toxic Chat (Lin et al., 2023), and LMSYS-Chat-1M (Zheng et al., 2023) by Li et al. (2024).

- **Toxic Chat T5:** guardrail model obtained by fine-tuning T5-Large (Raffel et al., 2020) on Toxic Chat (Lin et al., 2023).
- **ToxiGen HateBERT:** content moderation model obtained by fine-tuning HateBERT (Caselli et al., 2021) on ToxiGen (Hartvigsen et al., 2022).
- **ToxiGen RoBERTa:** content moderation model obtained by fine-tuning ToxDectRoBERTa (Zhou et al., 2021) on ToxiGen (Hartvigsen et al., 2022).
- **Detoxify Original:** BERT Base Uncased (Devlin et al., 2019) fine-tuned on Jigsaw's Toxic Comment Classification Challenge dataset (cjadams et al., 2019) for content moderation by Unitary AI (2020).
- **Detoxify Unbiased:** RoBERTa Base (Liu et al., 2019) fine-tuned on Jigsaw's Unintended Bias in Toxicity Classification dataset (cjadams et al., 2017) for content moderation by Unitary AI (2020).
- **Detoxify Multilingual:** XLM RoBERTa Base (Conneau et al., 2020) fine-tuned on Jigsaw's Multilingual Toxic Comment Classification dataset (Kivlichan et al., 2020) for content moderation by Unitary AI (2020).
- **Mistral-7B-Instruct v0.2:** general-purpose, instruction-tuned LLM proposed by Jiang et al. (2023). We instruct the model to check the input safety using the moderation prompt provided by its authors[8].
- **Mistral with refined policy:** Mistral-7B-Instruct v0.2 with the moderation policy of MD-Judge. More details in Section 5.4.

4.2 Evaluation Metrics

To evaluate the effectiveness of the considered models, we rely on F1 and Recall (when a dataset only comprises unsafe samples). Unlike previous works (Inan et al., 2023; Markov et al., 2023), we do not employ the Area Under the Precision-Recall Curve (AUPRC) as we found it overemphasizes models' Precision at the expense of Recall in the case of binary classification, thus hiding significant performance details. Moreover, F1 and Recall do not require classification probabilities as AUPRC, making them more convenient for comparing closed-weight models. We rely on Scikit-Learn (Pedregosa et al., 2011) to compute metric scores.

[8]https://docs.mistral.ai/capabilities/guardrailing

5 Results and Discussion

In this section, we present the results of our comparative evaluation. First, we discuss the models' effectiveness in assessing user prompts and human-AI conversations safety in Section 5.1 and Section 5.2, respectively. Then, in Section 5.3, we show preliminary results on non-English prompts. Finally, we evaluate the importance of content moderation policies in Section 5.4. Note that the results of Mistral with refined policy are considered only in Section 5.4. We refer the reader to Table 2 for the model aliases used in Table 3.

5.1 Prompts Moderation

In this section, we discuss the performance of the compared models at detecting unsafe user prompts, i.e., inputs containing or eliciting unsafe information. As shown in the first part of Table 3, guardrail models outperform content moderation models, suggesting the latter are not well-suited for prompt moderation. However, we highlight that the considered guardrail models have several times the parameters of the largest content moderation model, ToxiGen RoBERTa. Quite interestingly, Mistral, the general-purpose model we tested, often achieves better results than Llama Guard despite not being fine-tuned for detecting unsafe content in prompts and human-AI conversations. Overall, the best performing models are Llama Guard Defensive and MD-Judge, both of which surpass Llama Guard 2 in terms of performance, despite the latter is the most recent and advanced model. However, we observe that Llama Guard Defensive exhibits a potentially *exaggerated safety* behavior, given its relatively low F1 score on XSTest, which was proposed by Röttger et al. (2023) to evaluate such behavior. Due to the close performance of Llama Guard Defensive and MD-Judge, there is no clear answer to **RQ1**.

5.2 Conversations Moderation

In this section, we discuss the performance of the compared models at detecting user and LLM unsafe utterances in conversations. Results are presented in the second part of Table 3. Unlike prompts classification, content moderation models often perform closer to guardrail models when assessing

Model	Alias	Category	Base Model	Params	Architecture	Reference
Llama Guard	LG	Guardrail	Llama 2 7B	6.74 B	Decoder-only	Inan et al. (2023)
Llama Guard 2	LG-2	Guardrail	Llama 3 8B	8.03 B	Decoder-only	N/A
Llama Guard Defensive	LG-D	Guardrail	Llama 2 7B	6.74 B	Decoder-only	Ghosh et al. (2024)
Llama Guard Permissive	LG-P	Guardrail	Llama 2 7B	6.74 B	Decoder-only	Ghosh et al. (2024)
MD-Judge	MD-J	Guardrail	Mistral 7B	7.24 B	Decoder-only	Li et al. (2024)
Toxic Chat T5	TC-T5	Guardrail	T5 Large	0.74 B	Encod-Decod	N/A
ToxiGen HateBERT	TG-B	Moderation	BERT Base Uncased	0.11 B	Encoder-only	Hartvigsen et al. (2022)
ToxiGen RoBERTa	TG-R	Moderation	RoBERTa Large	0.36 B	Encoder-only	Hartvigsen et al. (2022)
Detoxify Original	DT-O	Moderation	BERT Base Uncased	0.11 B	Encoder-only	Unitary AI (2020)
Detoxify Unbiased	DT-U	Moderation	RoBERTa Base	0.12 B	Encoder-only	Unitary AI (2020)
Detoxify Multilingual	DT-M	Moderation	XLM RoBERTa Base	0.28 B	Encoder-only	Unitary AI (2020)
Mistral-7B-Instruct v0.2	Mis	General Purpose	Mistral 7B	7.24 B	Decoder-only	Jiang et al. (2023)
Mistral with refined policy	Mis+	General Purpose	Mistral 7B	7.24 B	Decoder-only	Section 5.4

Table 2: Benchmarked models. Alias indicates the shortened names used in other tables.

Dataset	Metric	LG	LG-2	LG-D	LG-P	MD-J	TC-T5	TG-B	TG-R	DT-O	DT-U	DT-M	Mis	Mis+
AdvBench Behaviors	Recall	0.837	0.963	0.990	0.931	0.987	0.842	0.550	0.117	0.019	0.012	0.012	0.948	**0.992** ↑ ‡
HarmBench Behaviors	Recall	0.478	**0.812**	0.684	0.569	0.675	0.300	0.341	0.059	0.028	0.016	0.031	0.516	0.622 ↑
I-CoNa	Recall	0.916	0.798	**0.978**	0.966	0.871	0.287	0.882	0.764	0.253	0.483	0.517	0.640	0.910 ↑ ‡
I-Controversial	Recall	0.900	0.625	**0.975**	0.900	0.900	0.225	0.550	0.450	0.025	0.125	0.125	0.300	0.875 ↑
I-MaliciousInstructions	Recall	0.780	0.860	0.950	0.850	0.950	0.660	0.510	0.240	0.050	0.080	0.070	0.750	**0.980** ↑ ‡
I-Physical-Safety	F1	0.147	0.507	0.526	0.295	0.243	0.076	**0.655**	0.113	0.179	0.076	0.076	0.226	0.458 ↑ ‡
MaliciousInstruct	Recall	0.820	0.890	**1.000**	0.920	0.990	0.730	0.280	0.000	0.000	0.000	0.000	0.980	0.990 ↑
MITRE	Recall	0.128	**0.867**	0.813	0.505	0.739	0.217	0.511	0.000	0.000	0.000	0.000	0.356	0.367
StrongREJECT Instructions	Recall	0.831	0.953	**0.986**	0.930	0.972	0.399	0.460	0.160	0.023	0.047	0.047	0.803	0.930 ↑
TDCRedTeaming	Recall	0.800	0.820	**1.000**	0.920	0.980	0.600	0.720	0.140	0.040	0.020	0.040	0.740	0.940 ↑
CatQA	Recall	0.798	0.936	**0.980**	0.893	0.944	0.511	0.176	0.018	0.007	0.018	0.016	0.978	0.945 ‡
Do Anything Now Questions	Recall	0.492	0.592	0.631	0.526	0.610	0.374	0.103	0.031	0.000	0.003	0.000	**0.810**	0.574
DoNotAnswer	Recall	0.321	0.442	0.496	0.399	**0.501**	0.224	0.249	0.100	0.028	0.034	0.048	0.435	0.460 ↑
HarmfulQ	Recall	0.890	0.875	**0.970**	0.930	0.945	0.665	0.290	0.055	0.010	0.000	0.010	0.930	0.965 ↑ ‡
HarmfulQA Questions	Recall	0.408	0.548	**0.780**	0.522	0.666	0.263	0.111	0.003	0.000	0.000	0.000	0.638	0.683 ↑ ‡
HEx-PHI	Recall	0.724	0.939	0.952	0.867	0.942	0.506	0.470	0.115	0.021	0.045	0.052	0.906	**0.958** ↑ ‡
XSTest	F1	0.819	**0.891**	0.783	0.812	0.858	0.632	0.373	0.233	0.186	0.287	0.424	0.826	0.878 ↑ ‡
AdvBench Strings	Recall	0.807	0.782	0.948	0.882	0.929	0.540	0.869	0.704	0.638	0.596	0.599	0.911	**0.949** ↑ ‡
DecodingTrust Stereotypes	Recall	0.875	0.780	**0.993**	0.944	0.957	0.211	0.977	0.900	0.589	0.655	0.668	0.572	0.765 ↑
DynaHate	F1	**0.804**	0.766	0.750	0.783	0.788	0.421	0.698	0.645	0.549	0.567	0.590	0.712	0.771 ↑
HateCheck	F1	0.942	**0.945**	0.877	0.909	0.921	0.562	0.853	0.833	0.757	0.761	0.803	0.879	0.909 ↑
Hatemoji Check	F1	0.862	0.788	0.873	**0.898**	0.869	0.376	0.791	0.607	0.669	0.575	0.642	0.780	0.853 ↑
SafeText	F1	0.143	**0.579**	0.504	0.294	0.425	0.085	0.417	0.052	0.154	0.078	0.097	0.487	**0.579** ↑ ‡
ToxiGen	F1	0.784	0.673	0.760	0.795	**0.821**	0.297	0.793	0.741	0.411	0.393	0.418	0.648	0.787 ↑
AART	Recall	0.825	0.843	**0.952**	0.891	0.879	0.745	0.483	0.122	0.019	0.037	0.054	0.815	0.898 ↑ ‡
OpenAI Moderation Dataset	F1	0.744	0.761	0.658	0.756	0.774	0.695	0.559	0.644	0.646	0.672	0.688	0.720	**0.779** ↑ ‡
SimpleSafetyTests	Recall	0.860	0.920	**1.000**	0.940	0.970	0.640	0.620	0.230	0.170	0.280	0.280	0.870	0.980 ↑ ‡
Toxic Chat	F1	0.561	0.422	0.577	0.678	0.816*	**0.822***	0.339	0.315	0.265	0.279	0.321	0.415	0.671 ↑
BeaverTails 330k	F1	0.686	0.755	0.778	0.755	**0.887***	0.448	0.643	0.245	0.173	0.216	0.236	0.696	0.740 ↑
UnsafeQA	F1	0.668	0.787	0.792	0.793	**0.842**	0.559	0.674	0.160	0.046	0.058	0.072	0.758	0.769 ↑
Bot-Adversarial Dialogue	F1	0.633	0.552	0.602	0.622	**0.652**	0.259	0.557	0.515	0.350	0.406	0.432	0.587	0.615 ↑
ConvAbuse	F1	0.000	0.348	0.663	0.676	0.704	0.575	0.427	0.625	0.669	0.674	0.676	0.582	**0.728** ↑ ‡
DICES 350	F1	0.270	0.182	0.327	0.298	**0.342**	0.142	0.316	0.200	0.075	0.103	0.124	0.276	0.225
DICES 990	F1	0.417	0.369	0.453	0.467	**0.555**	0.255	0.340	0.435	0.433	0.474	0.456	0.433	0.509 ↑
HarmfulQA	F1	0.171	0.391	**0.764**	0.563	0.676	0.204	0.565	0.000	0.000	0.000	0.000	0.648	0.427
ProsocialDialog	F1	0.519	0.383	**0.792**	0.691	0.720	0.337	0.689	0.471	0.371	0.389	0.411	0.697	0.762 ↑ ‡
PromptsEN (reference)	F1	0.816	0.828	0.850	0.841	**0.861**	0.583	0.651	0.497	0.427	0.420	0.456	0.804	0.856↑
PromptsDE	F1	0.718	0.728	**0.819**	0.791	0.683	0.251	0.607	0.131	0.201	0.128	0.079	0.704	0.743↑ ‡
PromptsFR	F1	0.714	0.734	**0.825**	0.800	0.672	0.356	0.235	0.101	0.106	0.085	0.435	0.697	0.734↑ ‡
PromptsIT	F1	0.708	0.732	**0.819**	0.794	0.664	0.230	0.093	0.137	0.161	0.163	0.429	0.659	0.720↑ ‡
PromptsES	F1	0.734	0.759	**0.832**	0.812	0.721	0.341	0.050	0.169	0.149	0.175	0.432	0.709	0.764↑ ‡

Table 3: Evaluation results. Best results are highlighted in boldface. Second-best results are underlined. The symbol * indicates a model was trained on the training set of the corresponding dataset. The symbols ↑ and ‡ in the last column indicate improvements over Mistral-7B-Instruct v0.2 (Mis) and MD-Judge (MD-J), respectively.

safety in conversations, probably thanks to the additional contextual information. These results suggest smaller models could achieve comparable results to current guardrail models if provided with a content moderation policy that gives further contextualization for the classification task. Again, Mistral shows better performance than Llama Guard. Overall, MD-Judge achieves the best performance among all the considered models, outperforming the more recent Llama Guard 2, Llama Guard Defensive, and Llama Guard Permissive. To answer **RQ2**, MD-Judge is the best-performing model at moderating conversations. However, there is still a large margin for improvements. Moreover, we found ToxiGen HateBERT to perform close to Llama Guard, despite having 70x less parameters. Therefore, performance-cost trade-offs of using multi-billion models as safety filters should be further investigated.

5.3 Multi-Lingual Capabilities

In this section, we discuss the out-of-the-box multi-lingual capabilities of the compared models. For reference, we report the performance of every model on a dataset built by merging all the English prompt datasets we translated, which we call PromptsEN. We highlight that none of the model received specific fine-tuning on multi-lingual datasets for safety classification other than Detoxify Multilingual. However, both the Llama-based models and the Mistral-based models were exposed to multi-lingual texts during pre-training. As shown in the third part of Table 3, Llama Guard Defensive, Llama Guard Permissive, and MD-Judge are the best performing models on the reference English dataset. However, Llama Guard Defensive and Llama Guard Permissive show much better performance than MD-Judge on German, French, Italian, and Spanish prompts. Although they still suffer from a performance degradation, it is far less noticeable than all the other considered models, especially in the case of Llama Guard Defensive. To answer **RQ3**, multi-lingual capabilities of most of the compared models are not comparable to those on English texts. However, we found the results achieved by Llama Guard Defensive to be encouraging for the detection of unsafe non-English text.

5.4 Policy Comparison

As introduced in Section 4.1, guardrail models are usually prompted with a content moderation policy and asked whether the input violates such a policy. In this section, we discuss the impact of the content moderation policy on the evaluation results. Specifically, we evaluate the performance of Mistral with the MD-Judge's policy. MD-Judge is based on Mistral and was fine-tuned on multiple safety datasets, such as BeaverTails330K (Ji et al., 2023), Toxic Chat (Lin et al., 2023), and LMSYS-Chat-1M (Zheng et al., 2023). With this experiment, we aim to assess whether their noticeable performance difference is due to the extensive fine-tuning received by MD-Judge or by their different content moderation policies. We highlight that the semantic content of the two policies presents significant overlaps. However, they are written and structured differently. The last column of Table 3 (Mis+) reports the performance of Mistral when prompted with MD-Judge's content moderation policy. Quite surprisingly, when prompted with MD-Judge's content moderation policy, Mistral show a very significant performance uplift, often outperforming MD-Judge and even reaching state-of-the-art results on multiple datasets. Such finding raise some concerns. First, comparisons with general-purpose LLMs are not present in recent publications on guardrail models (Inan et al., 2023; Ghosh et al., 2024). Secondly, the available training datasets for prompts and conversation safety classification may be insufficient to strongly improve over instruction-following models prompted for moderation. Moreover, prompt engineering (White et al., 2023) the content moderation policy could be crucial to improve over the state-of-the-art. Our analysis of **RQ4** reveals that content moderation policies significantly impact the effectiveness of guardrails models. Therefore, crafting well-written policies will be crucial for achieving improvements.

6 Conclusion and Future Work

In this work, we proposed GuardBench, a large-scale benchmark for evaluating guardrail models. GuardBench comprises 40 datasets for prompts and conversations safety evaluation. We included 35 datasets in English from previous works and five new datasets. Specifically, we built a new dataset for conversation safety evaluation by generating 22k answers to unsafe prompts from previous works. Moreover, we translated 31k English prompts to German, French, Italian, and Spanish, producing the first large-scale prompts safety datasets in those languages. To facilitate the adoption of GuardBench by the research community, we released a Python library offering a convenient evaluation pipeline. We also conducted the first large-scale evaluation of state-of-the-art guardrail models, showing that those models perform close to each other when identifying unsafe prompts, while we register more pronounced differences when used to moderate conversations. Finally, we showed general-purpose and instruction-following models can achieve competitive results when correctly prompted for safety moderation. In the future, we plan to extend GuardBench with an enhanced evaluation procedure to provide more structured results over the different categories of unsafe content. Safety classification of prompts and conversation utterances remains an open problem with considerable room for improvement. Advancements in this area are of utmost importance to safely deploy Large Language Models in high-risk and safety-critical domains, such as healthcare, education, and finance.

Limitations

While providing a valuable resource for guardrail models evaluation, our work has several limitations. Our benchmark scope is limited to the safe/unsafe binary classification task of prompts and conversation utterances. It does not cover multi-class and multi-label cases, although unsafe content may be classified in several, sometimes overlapping, categories of harm. Moreover, content that is unsafe for certain applications, such as finance, or belonging to specific unsafe categories may be missing from the datasets included in our benchmark. Several datasets included in our benchmark only have negative predictive power (Gardner et al., 2020), i.e. they only provide unsafe samples, as reported in Table 1. Thus, their usage should be limited to evaluating a model's weaknesses in recognizing unsafe content rather than characterizing generalizable strengths. Therefore, claims about model quality should not be overextended based solely on positive results on those datasets. We did not conduct any evaluation in which the models are required to follow, for example, a more permissive content moderation policy for a specific use case instead of the one provided by their authors or to adhere to a different view of safety. Finally, due to hardware constraints, we mainly investigated models up to a scale of 8 billion parameters. We also did not consider closed-weight and commercial moderation models such as OpenAI Moderation API and Perspective API.

Ethical Statement

This research aims to advance the development of Trustworthy Generative AI systems by contributing to the design of robust and effective guardrail models. Our large-scale benchmark, GuardBench, enables a comprehensive assessment of the performance of these critical AI safety components. We acknowledge that our research involves the usage and generation of unsafe content. The processing and inclusion of this content in GuardBench were necessary to evaluate the effectiveness of guardrail models in accurately identifying unsafe content. This research has received approval from the Joint Research Centre's (JRC) Ethical Review Board. In our commitment to contributing to AI safety, we make GuardBench available to the scientific community as open source software. We also share our novel datasets under a research-only license, providing access to them upon justified request. This approach ensures that the benefits of our research are accessible while mitigating potential risks and promoting responsible use.

References

Lora Aroyo, Alex S. Taylor, Mark Díaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. 2023. DICES dataset: Diversity in conversational AI evaluation for safety. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

David Baidoo-Anu and Leticia Owusu Ansah. 2023. Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. *SSRN Electronic Journal*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019*, pages 54–63. Association for Computational Linguistics.

Rishabh Bhardwaj, Do Duc Anh, and Soujanya Poria. 2024. Language models are homer simpson! safety re-alignment of fine-tuned language models through task arithmetic. *CoRR*, abs/2402.11746.

Rishabh Bhardwaj and Soujanya Poria. 2023. Red-teaming large language models using chain of utterances for safety-alignment. *CoRR*, abs/2308.09662.

Manish Bhatt, Sahana Chennabasappa, Yue Li, Cyrus Nikolaidis, Daniel Song, Shengye Wan, Faizan Ahmad, Cornelius Aschermann, Yaohui Chen, Dhaval Kapil, David Molnar, Spencer Whitman, and Joshua

Saxe. 2024. Cyberseceval 2: A wide-ranging cybersecurity evaluation suite for large language models. *CoRR*, abs/2404.13161.

Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. 2023. Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions. *CoRR*, abs/2309.07875.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Boyang Chen, Zongxiao Wu, and Ruoran Zhao. 2023. From fiction to fact: the growing role of generative ai in business and finance. *Journal of Chinese Economic and Business Studies*, 21(4):471–496.

cjadams, Daniel Borkan, inversion, Jeffrey Sorensen, Lucas Dixon, Lucy Vasserman, and nithum. 2019. Jigsaw unintended bias in toxicity classification.

cjadams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, nithum, and Will Cukierski. 2017. Toxic comment classification challenge.

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, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. Convabuse: Data, analysis, and benchmarks for nuanced detection in conversational AI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7388–7403. Association for Computational Linguistics.

Amanda Cercas Curry and Verena Rieser. 2018. #metoo alexa: How conversational systems respond to sexual harassment. In *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing, EthNLP@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5, 2018*, pages 7–14. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*, pages 512–515. AAAI Press.

Ona de Gibert, Naiara Pérez, Aitor García Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online, ALW@EMNLP 2018, Brussels, Belgium, October 31, 2018*, pages 11–20. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4536–4545. Association for Computational Linguistics.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 345–363. Association for Computational Linguistics.

David Esiobu, Xiaoqing Tan, Saghar Hosseini, Megan Ung, Yuchen Zhang, Jude Fernandes, Jane Dwivedi-Yu, Eleonora Presani, Adina Williams, and Eric Smith. 2023. ROBBIE: Robust bias evaluation of large generative language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3764–3814, Singapore. Association for Computational Linguistics.

Eve Fleisig, Aubrie Amstutz, Chad Atalla, Su Lin Blodgett, Hal Daumé III, Alexandra Olteanu, Emily Sheng, Dan Vann, and Hanna Wallach. 2023. FairPrism: Evaluating fairness-related harms in text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6231–6251, Toronto, Canada. Association for Computational Linguistics.

Catherine A. Gao, Frederick M. Howard, Nikolay S. Markov, Emma C. Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T. Pearson. 2023. Comparing scientific abstracts generated by chatgpt to real abstracts with detectors and blinded human reviewers. *npj Digital Medicine*, 6(1):75.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1307–1323. Association for Computational Linguistics.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 3356–3369. Association for Computational Linguistics.

Shaona Ghosh, Prasoon Varshney, Erick Galinkin, and Christopher Parisien. 2024. AEGIS: online adaptive AI content safety moderation with ensemble of LLM experts. *CoRR*, abs/2404.05993.

Lara Grimminger and Roman Klinger. 2021. Hate towards the political opponent: A twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EACL 2021, Online, April 19, 2021*, pages 171–180. Association for Computational Linguistics.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Z. Margetts. 2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1336–1350. Association for Computational Linguistics.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Steffen Herbold, Annette Hautli-Janisz, Ute Heuer, Zlata Kikteva, and Alexander Trautsch. 2023. A large-scale comparison of human-written versus chatgpt-generated essays. *Scientific Reports*, 13(1):18617.

Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. 2023. Catastrophic jailbreak of open-source llms via exploiting generation. *CoRR*, abs/2310.06987.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *CoRR*, abs/2312.06674.

Jiaming Ji, Mickel Liu, Josef 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 *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, Joe Hoover, Aida Azatian, Alyzeh Hussain, Austin Lara, Gabriel Cardenas, Adam Omary, Christina Park, Xin Wang, Clarisa Wijaya, Yong Zhang, Beth Meyerowitz, and Morteza Dehghani. 2022. Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale. *Lang. Resour. Evaluation*, 56(1):79–108.

Chris J. Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *CoRR*, abs/2009.10277.

Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. Prosocialdialog: A prosocial backbone for conversational agents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4005–4029. Association for Computational Linguistics.

Hannah Kirk, Bertie Vidgen, Paul Röttger, Tristan Thrush, and Scott A. Hale. 2022. Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1352–1368. Association for Computational Linguistics.

Ian Kivlichan, Jeffrey Sorensen, Julia Elliott, Lucy Vasserman, Martin Görner, and Phil Culliton. 2020. Jigsaw multilingual toxic comment classification.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A multilingual and document-level large audited dataset. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Nayeon Lee, Andrea Madotto, and Pascale Fung. 2019. Exploring social bias in chatbots using stereotype knowledge. In *Proceedings of the 2019 Workshop on Widening NLP@ACL 2019, Florence, Italy, July 28, 2019*, pages 177–180. Association for Computational Linguistics.

Sharon Levy, Emily Allaway, Melanie Subbiah, Lydia B. Chilton, Desmond Patton, Kathleen R. McKeown, and William Yang Wang. 2022. Safetext: A benchmark for exploring physical safety in language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2407–2421. Association for Computational Linguistics.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Sasko, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander M. Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2021, Online and Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 175–184. Association for Computational Linguistics.

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. *CoRR*, abs/2402.05044.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 140:1–55.

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, Singapore, December 6-10, 2023*, pages 4694–4702. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

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 *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 15009–15018. AAAI Press.

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. *CoRR*, abs/2402.04249.

Mantas Mazeika, Andy Zou, Norman Mu, Long Phan, Zifan Wang, Chunru Yu, Adam Khoja, Fengqing Jiang, Aidan O'Gara, Ellie Sakhaee, Zhen Xiang, Arezoo Rajabi, Dan Hendrycks, Radha Poovendran, Bo Li, and David Forsyth. 2023. Tdc 2023 (llm edition): The trojan detection challenge. In *NeurIPS Competition Track*.

Bertalan Meskó and Eric J. Topol. 2023. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *npj Digit. Medicine*, 6.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Junaid Qadir. 2023. Engineering education in the era of chatgpt: Promise and pitfalls of generative AI for education. In *IEEE Global Engineering Education Conference, EDUCON 2023, Kuwait, May 1-4, 2023*, pages 1–9. IEEE.

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! *CoRR*, abs/2310.03693.

Bhaktipriya Radharapu, Kevin Robinson, Lora Aroyo, and Preethi Lahoti. 2023. AART: ai-assisted red-teaming with diverse data generation for new llm-powered applications. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: EMNLP 2023 - Industry Track*, *Singapore, December 6-10, 2023*, pages 380–395. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2023. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. *CoRR*, abs/2308.01263.

Paul Röttger, Fabio Pernisi, Bertie Vidgen, and Dirk Hovy. 2024. Safetyprompts: a systematic review of open datasets for evaluating and improving large language model safety. *CoRR*, abs/2404.05399.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Z. Margetts, and Janet B. Pierrehumbert. 2021. Hatecheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 41–58. Association for Computational Linguistics.

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, ACL 2020, Online, July 5-10, 2020*, pages 5477–5490. Association for Computational Linguistics.

Omar Shaikh, Hongxin Zhang, William Held, Michael S. Bernstein, and Diyi Yang. 2023. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4454–4470. Association for Computational Linguistics.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2023. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *CoRR*, abs/2308.03825.

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. *CoRR*, abs/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, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Unitary AI. 2020. Detoxify. Github.

Bertie Vidgen, Adarsh Agrawal, Ahmed M. Ahmed, Victor Akinwande, Namir Al-Nuaimi, Najla Alfaraj, Elie Alhajjar, Lora Aroyo, Trupti Bavalatti, Borhane Blili-Hamelin, Kurt D. Bollacker, Rishi Bomassani, Marisa Ferrara Boston, Siméon Campos, Kal Chakra, Canyu Chen, Cody Coleman, Zacharie Delpierre Coudert, Leon Derczynski, Debojyoti Dutta, Ian Eisenberg, James Ezick, Heather Frase, Brian Fuller, Ram Gandikota, Agasthya Gangavarapu, Ananya Gangavarapu, James Gealy, Rajat Ghosh, James Goel, Usman Gohar, Subhra S. Goswami, Scott A. Hale, Wiebke Hutiri, Joseph Marvin Imperial, Surgan Jandial, Nick Judd, Felix Juefei-Xu, Foutse Khomh, Bhavya Kailkhura, Hannah Rose Kirk, Kevin Klyman, Chris Knotz, Michael Kuchnik, Shachi H. Kumar, Chris Lengerich, Bo Li, Zeyi Liao, Eileen Peters Long, Victor Lu, Yifan Mai, Priyanka Mary Mammen, Kelvin Manyeki, Sean McGregor, Virendra Mehta, Shafee Mohammed, Emanuel Moss, Lama Nachman, Dinesh Jinenhally Naganna, Amin Nikanjam, Besmira Nushi, Luis Oala, Iftach Orr, Alicia Parrish, Cigdem Patlak, William Pietri, Forough Poursabzi-Sangdeh, Eleonora Presani, Fabrizio Puletti, Paul Röttger, Saurav Sahay, Tim Santos, Nino Scherrer, Alice Schoenauer Sebag, Patrick Schramowski, Abolfazl Shahbazi, Vin Sharma, Xudong Shen, Vamsi Sistla, Leonard Tang, Davide Testuggine, Vithursan Thangarasa, Elizabeth Anne Watkins, Rebecca Weiss, Chris Welty, Tyler Wilbers, Adina Williams, Carole-Jean Wu, Poonam Yadav, Xianjun Yang, Yi Zeng, Wenhui Zhang, Fedor Zhdanov, Jiacheng Zhu, Percy Liang, Peter Mattson, and Joaquin Vanschoren. 2024. Introducing v0.5 of the AI safety benchmark from mlcommons. *CoRR*, abs/2404.12241.

Bertie Vidgen, Hannah Rose Kirk, Rebecca Qian, Nino Scherrer, Anand Kannappan, Scott A. Hale, and Paul Röttger. 2023. Simplesafetytests: a test suite for identifying critical safety risks in large language models. *CoRR*, abs/2311.08370.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1667–1682. Association for Computational Linguistics.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. 2023a. Decodingtrust: A comprehensive assessment of trustworthiness in GPT models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023b. Aligning large language models with human: A survey. *CoRR*, abs/2307.12966.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. 2024. Do-not-answer: Evaluating safeguards in llms. In *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, pages 896–911. Association for Computational Linguistics.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does LLM safety training fail? In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *CoRR*, abs/2302.11382.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2021. Bot-adversarial dialogue for safe conversational agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 2950–2968. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019*, pages 75–86. Association for Computational Linguistics.

Peng Zhang and Maged N. Kamel Boulos. 2023. Generative AI in medicine and healthcare: Promises, opportunities and challenges. *Future Internet*, 15(9):286.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2023. Safetybench: Evaluating the safety of large language models with multiple choice questions. *CoRR*, abs/2309.07045.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2023. Lmsys-chat-1m: A large-scale real-world LLM conversation dataset. *CoRR*, abs/2309.11998.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah Smith. 2021. Challenges in automated debiasing for toxic language detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3143–3155, Online. Association for Computational Linguistics.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *CoRR*, abs/2307.15043.

A Appendix

A.1 Labels Binarization

In this section, we provide further information on how we converted the labels of the gathered datasets into binary format. As BeaverTails 330k, ConvAbuse, DICES 350, and DICES 990 provide multiple annotations for each sample, we relied on a majority vote to decide whether a sample was safe or unsafe. We labelled samples as safe in case of ties. Note that some datasets use different binary labels for the safe and unsafe samples, such as toxic vs non-toxic. However, they directly fall within our definition of safe and unsafe content.

A.1.1 Prompts: Instructions

AdvBench Behaviors: Only unsafe samples. No conversion needed.

HarmBench Behaviors: Only unsafe samples. No conversion needed.

I-CoNa: Only unsafe samples. No conversion needed.

I-Controversial: Only unsafe samples. No conversion needed.

I-MaliciousInstructions: Only unsafe samples. No conversion needed.

I-Physical-Safety: Samples are labelled as safe or unsafe. No conversion needed.

MaliciousInstruct: Only unsafe samples. No conversion needed.
MITRE: Only unsafe samples. No conversion needed.
StrongREJECT Instructions: Only unsafe samples. No conversion needed.
TDCRedTeaming Instructions: Only unsafe samples. No conversion needed.

A.1.2 Prompts: Questions

CatQA: Only unsafe samples. No conversion needed.
Do Anything Now Questions: Only unsafe samples. No conversion needed.
DoNotAnswer: Only unsafe samples. No conversion needed.
HarmfulQ: Only unsafe samples. No conversion needed.
HarmfulQA Questions: Only unsafe samples. No conversion needed.
HEx-PHI: Only unsafe samples. No conversion needed.
XSTest: Samples are labelled as safe or unsafe. No conversion needed.

A.1.3 Prompts: Statements

AdvBench Strings: Only unsafe samples. No conversion needed.
DecodingTrust Stereotypes: Only unsafe samples. No conversion needed.
DynaHate: Samples are labelled as safe or unsafe. No conversion needed.
HateCheck: Samples are labelled as safe or unsafe. No conversion needed.
Hatemoji Check: Samples are labelled as safe or unsafe. No conversion needed.
SafeText: Samples are labelled as safe or unsafe. No conversion needed.
ToxiGen: Samples are labelled as safe or unsafe by following the authors instructions[9]. No conversion needed.

A.1.4 Prompts: Mixed

AART: Only unsafe samples. No conversion needed.
OpenAI Moderation Dataset: Samples have binary labels for each of the following categories: (S) Sexual, (H) Hate, (V) Violence, (HR) Harassment, (SH) Self-harm, (S3) Sexual/Minors, (H2) Hate/Threatening, and (V2) Violence/Graphic. We labelled each sample that does not belong to those categories as safe. Otherwise, we labelled it as unsafe.
SimpleSafetyTests: Samples are all unsafe and categorized as either (1) Suicide, Self-Harm, and Eating Disorders, (2) Physical Harm and Violence, (3) Illegal and Highly Regulated items, (4) Scams and Fraud, or (5) Child Abuse. We labelled all the samples as unsafe.
Toxic Chat: Samples are labelled as safe or unsafe. No conversion needed.

A.1.5 Conversations: Single-Turn

BeaverTails 330k: Samples have binary labels for each of the following categories: (1) Animal Abuse, (2) Child Abuse, (3) Controversial Topics, Politics, (4) Discrimination, Stereotype, Injustice, (5) Drug Abuse, Weapons, Banned Substance, (6) Financial Crime, Property Crime, Theft, (7) Hate Speech, Offensive Language, (8) Misinformation Regarding ethics, laws, and safety, (9) Non-Violent Unethical Behavior, (10) Privacy Violation, (11) Self-Harm, (12) Sexually Explicit, Adult Content, (13) Terrorism, Organized Crime, and (14) Violence, Aiding and Abetting, Incitement. We labelled each sample that does not belong to those categories as safe. Otherwise, we labelled it as unsafe.

A.1.6 Conversations: Multi-Turn

Bot-Adversarial Dialogue: Samples are labelled as safe or unsafe. No conversion needed.
ConvAbuse: Samples are labelled as safe or unsafe. No conversion needed.
DICES 350: Samples are labelled as safe or unsafe. No conversion needed.
DICES 990: Samples are labelled as safe or unsafe. No conversion needed.
HarmfulQA: Samples are labelled as safe or unsafe. No conversion needed.
ProsocialDialog: Samples are labelled as either (1) casual, (2) possibly needs caution, (3) probably needs caution, (4) needs caution, or (5) needs intervention. We converted (1), (2), and (3) as safe, and (4) and (5) as unsafe.

A.2 Novel Datasets

In this section, we provide further information regarding which datasets we translated (see Section 3.3) and those we answered to build UnsafeQA (see Section 3.4). Table 4 show which datasets were used as sources for deriving our datasets. In both cases, we excluded HEx-PHI (Qi et al., 2023) and

[9] `https://github.com/microsoft/TOXIGEN/blob/7b3495f467326cb91a69a9556b16be7ef0e06911/toxigen/utils.py#L9C1`

DecodingTrust Stereotypes (Wang et al., 2023a) due to their license limitations.

A.2.1 Multi-lingual Datasets

As shown in Table 4, we only considered prompt datasets for translation. We resorted to this choice as we found that the employed translation model MADLAD-400-3B-MT, which achieves state-of-the-art results for its size, performs well on prompts (usually short text in our benchmark) but struggles on responses due to their length. Thus, we only translated prompts. We did not include MITRE (Bhatt et al., 2024) in our derived datasets due to the inaccurate translations produced by MADLAD-400-3B-MT. This was likely caused by the highly technical nature of the prompts (cyber attacks) and their considerable length. For all the other prompts, the translations were favorably judged by our annotators, as reported in Section 3.3.

We highlight that we tested several translation models of different sizes before deciding to employ MADLAD-400-3B-MT. Specifically, we tested Google's MADLAD-400-7B-MT[10] and MADLAD-400-10B-MT[11] models, which showed the same issues on longer texts than the 3B variant we employed while producing comparable translations on short texts. We also tested Meta's SeamlessM4T v2[12] and Helsinki-NLP's OpusMT models[13][14][15][16], but we found them less reliable than MADLAD-400-3B-MT.

A.2.2 UnsafeQA

When building UnsafeQA, our goal was to provide the research community with a novel single-turn conversation dataset containing question-answer pairs. Thus, we answered questions and instructions with no available responses, as reported in Table 4. Instructions are requests phrased as commands, which are functionally similar to questions. Thus, we included them. For instance, the question *How can I make a bomb?* can be rephrased as the following instruction: *Guide me in making a bomb.* We excluded statement datasets from UnsafeQA as they are functionally different from questions. We included mixed datasets as they mostly contain questions and instructions.

[10] https://huggingface.co/google/madlad400-7b-mt

[11] https://huggingface.co/google/madlad400-10b-mt

[12] https://huggingface.co/facebook/seamless-m4t-v2-large

[13] https://huggingface.co/Helsinki-NLP/opus-mt-en-de

[14] https://huggingface.co/Helsinki-NLP/opus-mt-en-fr

[15] https://huggingface.co/Helsinki-NLP/opus-mt-en-it

[16] https://huggingface.co/Helsinki-NLP/opus-mt-en-es

Dataset	Category	Sub-category	License	Reference	PromptsDE	PromptsFR	PromptsIT	PromptsES	UnsafeQA
AdvBench Behaviors	Prompts	Instructions	MIT	Zou et al. (2023)	✓	✓	✓	✓	✓
HarmBench Behaviors	Prompts	Instructions	MIT	Mazeika et al. (2024)	✓	✓	✓	✓	✓
I-CoNa	Prompts	Instructions	CC BY-NC 4.0	Bianchi et al. (2023)	✓	✓	✓	✓	✓
I-Controversial	Prompts	Instructions	CC BY-NC 4.0	Bianchi et al. (2023)	✓	✓	✓	✓	✓
I-MaliciousInstructions	Prompts	Instructions	CC BY-NC 4.0	Bianchi et al. (2023)	✓	✓	✓	✓	✓
I-Physical-Safety	Prompts	Instructions	CC BY-NC 4.0	Bianchi et al. (2023)	✓	✓	✓	✓	✓
MaliciousInstruct	Prompts	Instructions	MIT	Huang et al. (2023)	✓	✓	✓	✓	✓
MITRE	Prompts	Instructions	MIT	Bhatt et al. (2024)	✗	✗	✗	✗	✓
StrongREJECT Instructions	Prompts	Instructions	MIT	Souly et al. (2024)	✓	✓	✓	✓	✓
TDCRedTeaming Instructions	Prompts	Instructions	MIT	Mazeika et al. (2023)	✓	✓	✓	✓	✓
CatQA	Prompts	Questions	Apache 2.0	Bhardwaj et al. (2024)	✓	✓	✓	✓	✓
Do Anything Now Questions	Prompts	Questions	MIT	Shen et al. (2023)	✓	✓	✓	✓	✓
DoNotAnswer	Prompts	Questions	Apache 2.0	Wang et al. (2024)	✓	✓	✓	✓	✓
HarmfulQ	Prompts	Questions	MIT	Shaikh et al. (2023)	✓	✓	✓	✓	✓
HarmfulQA Questions	Prompts	Questions	Apache 2.0	Bhardwaj and Poria (2023)	✓	✓	✓	✓	✓
HEx-PHI	Prompts	Questions	Custom	Qi et al. (2023)	✗	✗	✗	✗	✗
XSTest	Prompts	Questions	CC BY 4.0	Röttger et al. (2023)	✓	✓	✓	✓	✓
AdvBench Strings	Prompts	Statements	MIT	Zou et al. (2023)	✓	✓	✓	✓	✗
DecodingTrust Stereotypes	Prompts	Statements	CC BY-SA 4.0	Wang et al. (2023a)	✗	✗	✗	✗	✗
DynaHate	Prompts	Statements	Apache 2.0	Vidgen et al. (2021)	✓	✓	✓	✓	✗
HateCheck	Prompts	Statements	CC BY 4.0	Röttger et al. (2021)	✓	✓	✓	✓	✗
Hatemoji Check	Prompts	Statements	CC BY 4.0	Kirk et al. (2022)	✓	✓	✓	✓	✗
SafeText	Prompts	Statements	MIT	Levy et al. (2022)	✓	✓	✓	✓	✗
ToxiGen	Prompts	Statements	MIT	Hartvigsen et al. (2022)	✓	✓	✓	✓	✗
AART	Prompts	Mixed	CC BY 4.0	Radharapu et al. (2023)	✓	✓	✓	✓	✓
OpenAI Moderation Dataset	Prompts	Mixed	MIT	Markov et al. (2023)	✓	✓	✓	✓	✓
SimpleSafetyTests	Prompts	Mixed	CC BY 4.0	Vidgen et al. (2023)	✓	✓	✓	✓	✓
Toxic Chat	Prompts	Mixed	CC BY-NC 4.0	Lin et al. (2023)	✓	✓	✓	✓	✓
BeaverTails 330k	Conversations	Single-Turn	MIT	Ji et al. (2023)	✗	✗	✗	✗	✗
Bot-Adversarial Dialogue	Conversations	Multi-Turn	Apache 2.0	Xu et al. (2021)	✗	✗	✗	✗	✗
ConvAbuse	Conversations	Multi-Turn	CC BY 4.0	Curry et al. (2021)	✗	✗	✗	✗	✗
DICES 350	Conversations	Multi-Turn	CC BY 4.0	Aroyo et al. (2023)	✗	✗	✗	✗	✗
DICES 990	Conversations	Multi-Turn	CC BY 4.0	Aroyo et al. (2023)	✗	✗	✗	✗	✗
HarmfulQA	Conversations	Multi-Turn	Apache 2.0	Bhardwaj and Poria (2023)	✗	✗	✗	✗	✗
ProsocialDialog	Conversations	Multi-Turn	CC BY 4.0	Kim et al. (2022)	✗	✗	✗	✗	✗

Table 4: Datasets used to derive our multi-lingual datasets and Unsafe QA.