2/9/2026 1:08:18 PM

Compare Results

Old File:

2024.emnlp-main.1022.pdf

17 pages (16.35 MB)

versus

New File:

2024_emnlp-main_1022.pdf

13 pages (299 KB)

2/8/2026 5:27:39 AM

Total Changes

45

Content

19 Replacements

11 Insertions

15 Deletions

Styling and Annotations

0 Styling

0 Annotations

Go to First Change (page 1)

](https://github.com/AmenRa/guardbench)

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.europ

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 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 (Mesk'o 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 re-

*[https://github.com/AmenRa/guardbench

sponses 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 multilingual 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 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

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; Gimminger 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 (R"ottger 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 (R"ottger 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 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 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 (R"ottger 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, OpenAI Moderation API), 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 our 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 R"ottger 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-400-38-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.

3.4 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. 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 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 repro-

duced 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 `.tex`, 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.
- 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) 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).

[ILLEGIBLE]

Table 1: Table 1: List of benchmark datasets. [ILLEGIBLE]

- 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.
- 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.

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

[ILLEGIBLE]

Table 2: Table 2: Benchmarked models. [ILLEGIBLE]

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 R"ottger 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 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

[ILLEGIBLE]

Table 3: Table 3: Evaluation results. [ILLEGIBLE]

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

[1] *[ILLEGIBLE if any characters appear unclear below; line breaks preserved as in source.]*[0.5em] **References**

Lora Aroyo, Alex S. Taylor, Mark Diaz, Christophe
Homan, Alicia Panish, Gregory Serapio-Garcia,
Vinodkumar Prabhakaran, and Ding Wang. 2023.
DICES dataset: Diversity in conversational AI eva
Processing Systems 36: Annual Conference on Neura
2023, New Orleans, I-4, USA, December l0 - 16,
2023.

David Baidoo-Anu and Leticia Owusu Ansah. 2023. E
(ai): Understanding the potential benefits ofchat
in promoting teaching and learning. SSRN Electron
Journal.

Valerio Basile, Cristina Bosco, Elisabetta Fersin
Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019
Semeval-2019 task 5: Multilingual detection of ha
speech against immigrants and women in twitter. I
Proceedings of the l3th 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 throug
task arithmetic. CoRR, abs/2402.11746.

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

[ILLEGIBLE]

A Labels Binarization

A.1 LabelsBinarization

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 per data point, we label a sampl as unsafe if at least one of the annotators labeled i unsafe. For all other datasets, we map the available to safe/unsafe. [ILLEGIBLE]

A.2 Novel Datasets

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 to derive the two novel datasets and whether we answered and/or translated them. [ILLEGIBLE]

[ILLEGIBLE]

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