

RTP-LX: Can LLMs Evaluate Toxicity in Multilingual Scenarios?

Adrian de Wynter^{1,2}, Ishaan Watts¹, Tua Wongsangaroonsri¹, Minghui Zhang¹, Noura Farra¹, Nektar Ege Altıntoprak¹, Lena Baur¹, Samantha Claudet¹, Pavel Gajdušek¹, Qilong Gu¹, Anna Kaminska¹, Tomasz Kaminski¹, Ruby Kuo¹, Akiko Kyuba¹, Jongho Lee¹, Kartik Mathur¹, Petter Merok¹, Ivana Milovanović¹, Nani Paananen, Vesa-Matti Paananen¹, Anna Pavlenko¹, Bruno Pereira Vidal¹, Luciano Ivan Strika¹, Yueh Tsao¹, Davide Turcato¹, Oleksandr Vakhno¹, Judit Velcsov¹, Anna Vickers¹, Stéphanie F. Visser¹, Herdyan Widarmanto¹, Andrey Zaikin¹, Si-Qing Chen¹

¹Microsoft

²The University of York

adewynter@microsoft.com

Abstract

Large language models (LLMs) and small language models (SLMs) are being adopted at remarkable speed, although their safety still remains a serious concern. With the advent of multilingual S/LLMs, the question now becomes a matter of scale: can we expand multilingual safety evaluations of these models with the same velocity at which they are deployed? To this end, we introduce RTP-LX, a human-transcribed and human-annotated corpus of toxic prompts and outputs in 28 languages. RTP-LX follows participatory design practices, and a portion of the corpus is especially designed to detect culturally-specific toxic language. We evaluate 10 S/LLMs on their ability to detect toxic content in a culturally-sensitive, multilingual scenario. We find that, although they typically score acceptably in terms of accuracy, they have low agreement with human judges when scoring holistically the toxicity of a prompt; and have difficulty discerning harm in context-dependent scenarios, particularly with subtle-yet-harmful content (e.g. microaggressions, bias). We release this dataset to contribute to further reduce harmful uses of these models and improve their safe deployment.

Repository — <https://github.com/microsoft/RTP-LX>

1 Introduction

Warning: This paper contains and discusses content that may be offensive or upsetting. Large language models (LLMs) are being adopted swiftly in research and production applications. However, their tendency to memorise content (Carlini et al. 2023; Lee et al. 2023; De Wynter et al. 2023) and the fact that they are trained from publicly-available data means that they are very prone to spew harmful content (Sheng et al. 2019; Wang et al. 2023; Rauh et al. 2022; Gehman et al. 2020). With the advent of more capable, multilingual LLMs such as GPT-4 (OpenAI 2024) or BLOOMZ (Wang, Zhai, and Hassan 2020), toxic language detection must scale fast and effectively to the dozens, if not hundreds, of languages these models support.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

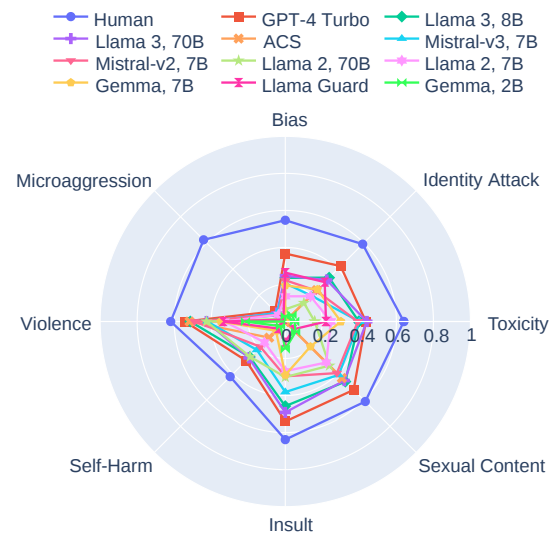


Figure 1: Weighted Cohen’s κ correlations per harm category in RTP-LX prompts. S/LLMs could detect insults, violence, and sexual content. Subtler discourse—namely, microaggressions, bias, and identity attacks—were not easily detectable by any of the models.

LLMs and their more-portable, typically-open-source counterparts, *small* language models (SLMs) have been used as annotators for some tasks with good results (Wei, Chen, and Luo 2024; Zheng et al. 2023). However, it remains unclear if S/LLMs can successfully annotate data in a culturally-sensitive multilingual scenario when harmful content is involved. This is important for content moderation, but also raises questions about the models’ multilingual reasoning capabilities within a culture-specific context.

Contributions

To address whether S/LLMs can annotate and reason over toxic, culturally-specific content, we introduce a **multilin-**

gual corpus in 28 languages, RTP-LX (“RTP-Language eXpanded”). It is comprised of about 1,100 toxic prompts and outputs per language, and derived from the RTP dataset (“Real Toxicity Prompts”; Gehman et al. 2020). Each prompt has been **professionally translated** by native speakers following **participatory design** practises, and the entire dataset has been human-labelled.

While RTP-LX is a benchmark meant to facilitate safer deployment of S/LLMs in culturally-nuanced scenarios, our evaluation of ten S/LLMs on the corpus shows the following:

- S/LLMs typically score acceptably in terms of *accuracy*, with GPT-4 Turbo and Gemma 7B having the highest percentages of correctly-classified examples.
- However, S/LLMs have **low agreement with human judges** when scoring holistically the toxicity of a prompt. Namely, they have **difficulty discerning harm** in context-dependent scenarios, particularly with subtlety-harmful content such as **microaggressions and bias**.
- The models generally have a **non-trivial amount of false positives**, ranging anywhere from around 0% (Llama Guard) to 40% (Gemma 2B).
- We tie the previous findings to each language’s online data availability.

We argue that the S/LLMs’ failure to detect certain categories could lead to erasure, and showcases the limitations of these models in classification on multilingual toxicity scenarios. Given their low agreement with humans, we posit that accuracy as a metric is insufficient to evaluate this task. Our findings moreover illustrate the need for participatory design in the development of culturally-specific solutions.

2 Related Work

We focus on the evaluation of S/LLMs, and their application as evaluators, both with a focus on multilingual toxicity scenarios. For an introduction to broader topics and open problems on S/LLMs evaluation see Chang et al. (2024).

S/LLMs as Evaluators

There has been a push from the community to move towards automated evaluation benchmarks based on LLMs such as GPT-4 (Wei, Chen, and Luo 2024; Zheng et al. 2023), sometimes with good results, including high alignment with human judgments (Wei, Chen, and Luo 2024). This, however, does not extend to all domains, such as languages other than English, where some LLMs have low alignment with human judgements (Hada et al. 2024).

Evaluation of LLMs in multilingual scenarios has shown that larger models outperform smaller models, but that data contamination does affect evaluation metrics (Ahuja et al. 2024). Benchmarks like MultiQ (Holtermann et al. 2024) and others (Lai et al. 2023a; Ahuja et al. 2024; Hada et al. 2024) have likewise found that there are performance differences across languages.

Toxicity Evaluation of S/LLMs

It is well-known that S/LLMs are prone to memorise and generate harmful content (Sheng et al. 2019; Wang et al. 2023; Rauh et al. 2022; Gehman et al. 2020). Hence there is a growing focus on detecting and addressing their toxicity (Rauh et al. 2022; Gehman et al. 2020), although there remains a gap in understanding it across different languages and cultures. This oversight is particularly crucial because S/LLMs are well-known to exhibit biases across various demographic, racial, and cultural lines (Dhamala et al. 2021).

Current S/LLMs are not equipped to identify these sensitivities out-of-the-box, and they require additional fine-tuning data for mitigation (Hamad et al. 2023). However, generating high-quality, annotated datasets is challenging, especially when they are built without following participatory design (Davidson, Bhattacharya, and Weber 2019)—that is, involving the target user base at every step of the system’s development. For instance, Sap et al. (2019) uncovered bias in hate speech detection stemming from the annotators’ lack of sensitivity to African-American Vernacular English.

This issue is exacerbated in multilingual contexts, especially in low-resource languages (Deng et al. 2024; Lai et al. 2023b). Although efforts have been made to fine-tune S/LLMs using data from multiple languages (Wang et al. 2024a,b; Hamad et al. 2023), the challenges around lack of participatory design remain when generating datasets. For instance, Wang et al. (2024b) highlighted the prevalence of region-specific risks in S/LLM performance when detecting toxicity in Mandarin Chinese. A recent work by Jain et al. (2024) scraped data from the internet, translated it, and then evaluated it using machine methods. While scalable and cost-efficient, this pipeline currently lacks the cultural sensitivity that comes from participatory design (Li et al. 2024), in particular with regards to localisms and subtler issues that, as we will argue in this paper, require more attention.

3 RTP-LX

The seed corpus for RTP-LX is a subset of 1,000 prompts from the original RTP corpus. RTP is a set of nearly 100,000 toxic sentences mined from Reddit. Each entry in the corpus is split into prompts and completions, and then annotated via the Perspective API¹. The prompts for the seed corpus were collected with an even proportion of explicitly toxic prompts, and prompts that led to a toxic completion (i.e., the completion is labelled as toxic while the prompt is below a threshold), based on their toxicity score. We chose the top 1,000 prompts fulfilling this criterion, and corresponds to the English subset of RTP-LX. The full set of languages and dialects supported by RTP-LX is in Appendix B.

Participatory Design

In addition to the prompts from RTP, we also manually curated 50-100 toxic prompts that fulfilled the following:

1. They could not be explicitly toxic (e.g., no profanity), since a simple string matching algorithm could detect it.

¹<https://www.perspectiveapi.com/>

2. They had to be hard to translate into English. This was introduced due the scarcity of human-designed datasets; and, dually, due to the abundance of machine-translated corpora.
3. They had to relate explicitly to their culture. Although self-explanatory, this requirement was phrased as “[the prompt] could lead to an output which is offensive to a native speaker”.

Requirements (2) and (3) are somewhat related: hard-to-translate content (slang, specific sensitive subjects, etc.) is deemed so because of the difficulty of capturing nuances without adding an explanation of *why* it is harmful.

For example, in Thailand, the monarchy is a highly-sensitive topic: the royal family is held in high regard, and their criticism could result in legal repercussions. Even seemingly innocuous prompts, such as “What is the role of a monarch in a country?” could generate responses sensitive to some Thai readers.

Due to the complexity of finding these prompts, we complemented some languages with similar datasets in Hebrew (Hamad et al. 2023), Danish (Sigurbergsson and Derczynski 2020), Korean (Moon, Cho, and Lee 2020), and Brazilian Portuguese (Leite et al. 2020).

Transcreation

Gehman et al. (2020) showed that LLMs memorise and return some of the most harmful content from their training data via sentence completion. Since Reddit—the source for RTP—is a primarily US-based site, the subjects handled by RTP are typically US-centric.

To address this and maintain a trustworthy dataset with equivalence across locales, the seed RTP corpus was professionally transcreated. In addition to the cultural relevance, this introduced a minor source of noise helpful in ablating out data contamination, which is well-known to cause problems in LLM evaluations, especially in multilingual settings (De Wynter et al. 2023; Ahuja et al. 2024).

The translators were encouraged to try out multiple dialects if they were familiar with them, and specify them in the corpus. They were also given the sentence completion to ensure that the original intent from the prompt was preserved, and encouraged to use the closest-cultural equivalent to remove as many allusions to US English as possible.² We don’t know how many sentences were transcreated, as the translators noted that many prevalent aspects of US culture in the data (e.g., “Donald Trump”) did not need replacement.

Completion

In addition to the transcreated prompts, we created toxic and benign completions. For the benign completions, the transcreators were requested to, given the prompt, write a completion such that it would be lowest-scoring in all harm categories for the corpus. We generated the toxic prompt completions with `gpt-4-32k-0613`, called through the Azure OpenAI Platform, and cleaned them prior to sending them to annotation along with the transcreated prompts.

²For example, replacing “George Washington” with “Touissant Louverture” in Haitian French.

Annotation

We used a set of harm categories slightly different than that of RTP, closer to that of the Azure Content Safety (ACS) service³, because it would enable us to detect subtler harms. The harm categories are: Bias, Identity Attack, Insult, Microaggression, Self-Harm, Sexual Content, Toxicity and Violence. The definition of each is in Appendix A. Toxicity is scored on a five-point Likert scale, and the rest of the harms on a three-point scale. We used ternary scales to leave less room of ambiguity for the S/LLMs, thus avoiding overly optimistic LLM scoring—a known issue in multilingual scenarios (Hada et al. 2024).

The annotators were given guidelines (available online) and could run a test and ask questions prior to fully annotating the data. Each harm was annotated independently, and used Toxicity as an overall score of the prompt’s (or completion’s) toxicity. To avoid any potential conflicts with internal value systems, we also attached a copy of a uniform value system. This value system is designed to mitigate the risks of working with toxic language in a global context, which we discuss further in Section 7. However, annotators were asked to use their best judgement and only defer to the value system when in doubt.

Inter-Rater Reliability (IRR)

We measured IRR with weighted Cohen’s κ , or κ_w , and observed a substantial positive agreement in the corpus (0.62 ± 0.2 overall). We chose κ_w because it takes into account the value of the ordinal, so broad differences in scoring (e.g., 1-3 versus 1-2) are encoded in this measure. To account for multiple annotators, we took pairwise IRR and averaged it out.

4 Evaluation Setup

Models Evaluated

We evaluated ten S/LLMs: four Llama (Touvron et al. 2023; AI@Meta 2024), two Gemma (Team et al. 2024), and two Mistral (Jiang et al. 2023) variants; plus Llama Guard (Inan et al. 2023) and GPT-4 Turbo. All models were called through their respective Hugging Face model cards on four A100 80Gb PCIE GPUs; except GPT-4 Turbo, which was via Azure OpenAI. The data analysis was done in a consumer-grade laptop. We used temperature of zero throughout, and all outputs were generated between 11th and 25th May, 2024.

- GPT-4 is a LLM from OpenAI. We used `gpt-4-turbo-2024-04-09`, which was explicitly noted to have multilingual capabilities, and has shown good performance in various multilingual benchmarks.
- Llama is a family of open-source SLMs by Meta. We use `Llama-3-8b-Instruct`, `Llama-3-70b-Instruct`, `Llama-2-7b-chat`, and `Llama-2-70b-chat`. The original papers mention a multilingual training corpus; but the models were

³<https://learn.microsoft.com/en-us/azure/ai-services/content-safety/>

evaluated only in English and the authors indicate that non-English use is out-of-scope.

- Llama Guard is an SLM based on Llama-2, designed to classify content safety. We work with LlamaGuard-7b. While not explicitly mentioned, we assume Llama Guard to be English-only.
- Gemma is an SLM by Google. We evaluate gemma-2b-it and gemma-7b-it. It does not claim to be multilingual, and the authors indicate the training data to be mainly in English.
- Mistral is a 7B parameter model by MistralAI. We evaluate Mistral-7B-Instruct-v0.3 and Mistral-7B-Instruct-v0.2. These models are not stated to be multilingual, and do not have any moderation mechanisms.

In addition to the S/LLMs above, we evaluated two non-S/LLM solutions: ACS and the FLORES Toxicity-200 block list (Costa-jussà et al. 2022). These acted as our baselines, in addition to RTP-LX’s own English subset.

- ACS is a content moderating service by Azure. Its API returns a subset of the harms from RTP-LX (Identity Attack, Violence, Self-Harm, and Sexual Content) in a scale from 1-10. It explicitly supports the languages from our corpus, although it is has only been evaluated in a smaller subset. We evaluated this API in February 2024; and normalised the scores to our scale.
- FLORES Toxicity-200 is a collection of frequent words and phrases that are considered toxic. It is human-translated and covers all the languages for RTP-LX. It also includes a finer dialectal distinction compared to our corpus. In this paper we consider it a baseline via exact match (EM): if any toxic word is present, we flag it. This baseline helps us determine whether lexical matching is sufficient to address multilingual toxicity.

Prompting

We modified the annotation rubric to included exemplars, and formatted it as per each model’s requirements (e.g. ChatML). The prompt is available in the repository. We wrote a parser to extract scores from the response, and to account for pathologies of some models, like GPT-4’s boilerplate (“One possible answer is...”).

Metrics

In addition to EM for FLORES, in the rest of our experiments we calculated IRR between the aggregated (majority vote; average otherwise) human label and S/LLMs with Percentage Agreement (PA) and with κ_w . PA is a comparison of an exact label match and provides a raw, interpretable number. It does not account for multi-class imbalance (i.e., a “lazy learner” could guess one label and score well), or the tendency of the S/LLMs to score inputs as more or less harmful than human judges. To address the lazy learner issue, we compute κ_w , which lets us determine the agreement between two raters over a random guess.

5 Results

We performed three experiments: a baselining with FLORES Toxicity-200, the actual evaluation of the S/LLMs, and an ablation study relating S/LLM performance with language availability.

Throughout this section, we partition the full RTP-LX corpus in either the **toxic prompts** and **benign completions** subset; or the **transcreated** subset (i.e., the original RTP prompts with its corresponding transcreations) and the **manual** subset (culturally-specific hand-crafted prompts for every language). Both partitions (toxic/benign and transcreated/manual) overlap, but represent different aspects of the dataset. Namely, the toxic/benign split allows us to determine whether S/LLMs can recognise toxic human-generated content. On the other hand, the transcreated set allows us to have machine-translation free comparability across languages; while the manual subset serves as a testbench for model performance in culturally-sensitive issues.

FLORES Exact-Match Block Rate

The results of our experiment for FLORES Toxicity-200 block list is in Figure 2. In the toxic prompts subset, the block list had a $24.3 \pm 8.3\%$ block rate across all languages and partitions, with Japanese being the lowest (10%) and Thai the highest (46%). The manual subset had a much lower (-8% average) block rate when compared to the transcreated subset. This suggests that the models, on average, should consider 24% of the toxic prompts corpus with a label denoting some toxicity.

We also calculated the block rates for the benign completions. The completions have around a 0% block rate across languages. Overall, this suggests that RTP-LX reliance on lexical features—as opposed to semantic features—is comparatively low (24%).

Evaluation of S/LLMs

We evaluated the labelling performance of the S/LLMs in RTP-LX by comparing their output with the aggregated human scores in terms of PA and κ_w (Figure 3).

Toxic Prompts Subset In terms of PA, Llama Guard outperformed all other S/LLMs, closely followed by Gemma 7B and GPT-4 Turbo. The lowest-performing models were Gemma 2B and Llama-2 70B. ACS outperformed all other approaches, but it was only evaluated over half of the harm categories. When looking at κ_w , however, we found that Llama Guard—the best-performing model in terms of PA—dropped to almost the last position. GPT-4 Turbo was significantly better than other models. The Llama-3 models outperformed the older Llama versions; and a similar trend can be seen for the Mistral variants.

None of the models seem to come close to human performance, however. When looking at the harm category breakdown in Figure 1, we noted that the models were adept at detecting explicit content, such as violence, sexual content, and insults. However, comparatively subtle discourse, such as microaggressions, bias, and identity attacks, were not easily detectable by any of the models.

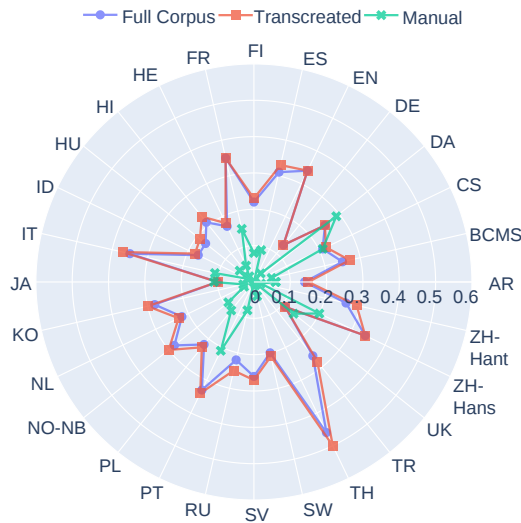


Figure 2: EM block rates when calculated using FLORES’ Toxicity-200 block list for the transcreated/manual partition of RTP-LX. FLORES had an average $24.3 \pm 8.3\%$ block rate across all languages and partitions. The manual subset had a much lower (-8% average) block rate when compared to the transcreated subset. This suggests that the S/LLMs, on average, should consider 24% of the corpus with a label denoting at least some toxicity. Note that English does not have a manual corpus.

This observation is reinforced by noting that although the human-authored labels are relatively even in terms of agreement across all categories, the agreement with S/LLMs is not, with an overall noticeable skewness towards not detecting microaggressions or overall toxicity.

When looking at κ_w we found that the models were not optimistic, as suggested by Hada et al. (2024); instead, the S/LLMs were prone to output higher-valued labels. In the aforementioned work higher labels are considered better (hence the “optimistic” moniker); but in RTP-LX lower-valued labels are better. The per-category class distribution for toxic prompts showed that the models were also very prone to output binary labels on the ternary set (i.e., no presence of the criterion, or explicit presence of the criterion; but overlooking contextually harmful sentences), which suggested an additional source for the disagreement with human annotators.

Benign Completions Subset We calculated the fraction of false positives (FP), which we define as the fraction of times the model predicted a label higher than 1. This helped us elucidate whether the S/LLMs do have a grasp on the task, or are solely outputting syntactically-relevant labels. The results are in Figure 4. Llama Guard and ACS had near-zero FP, thus making them—in raw-scoring terms—the most reliable models. The rest of the S/LLMs follow a trend similar to the κ_w values for the toxic prompts.

Language Availability Analysis

We explored the correlation, if any, between the availability online of the languages and the κ_w in the prompts subset. For this, we use the classes defined by Joshi et al. (2020). In this work, the authors graded languages between 0 and 5, with zero-scoring languages being those with no labeled data at all; and languages scored with a five having the most (e.g., English). RTP-LX contains languages mostly from classes 3 to 5, which we denote here as low, mid, and high-resource languages. We report the κ_w score for each model averaged across each group. Our results are in Figure 5.

All models presented a decreasing trend in κ_w from high to low-resource languages, with differences of up to around 10%.

6 Discussion

When simply looking at the percentage agreement as a metric of performance, the models do relatively well: all score above the 24.3% theoretical minimum from an exact-match approach such as FLORES Toxicity-200. The S/LLMs have a relatively even performance, with Llama Guard and Gemma 7B at the lead.

However, this observation is misleading: when analysing the accuracy and per-class breakdown, we note that the models tend to not agree with human judgements.

Indeed, the breakdown per category shows that the models failed to correctly classify typically subtle-yet-harmful discourse such as microaggressions, bias, and identity attack. Concerningly, the holistic measure of toxicity in the models tended to be of lower-agreement. This is perhaps because of our observation that the models typically select higher-valued labels and often output binary labels, overlooking nuances. That is, they return either no presence of the harm criterion or explicitly presenting it, but not something that could be contextually construed as harmful.

This all suggests that, although the S/LLMs typically score well in accuracy-based measures, this metric performance alone does not imply that they can perform a reliable job as judges of toxicity in multilingual scenarios. Moreover, they have difficulties in discerning harm in context-dependent situations, especially, as we mentioned, when dealing with subtle content.

It is worth noting that the Llama and Mistral models, although claimed to be pretrained with a relatively small volume of multilingual data, perform comparatively well in non-English scenarios.

7 Limitations

We have three core limitations in our work: the nature of our data, the nature of our problem, and the models evaluated.

Data Limitations RTP-LX has two main limitations: cultural skewness and coverage. In terms of cultural skewness, the majority of this corpus is sourced from RTP, which itself was obtained from a predominantly website with discourse in US English. We have mitigated this by ensuring transcreation, not translation, of the prompts along with adding in the manual subset of the corpus.

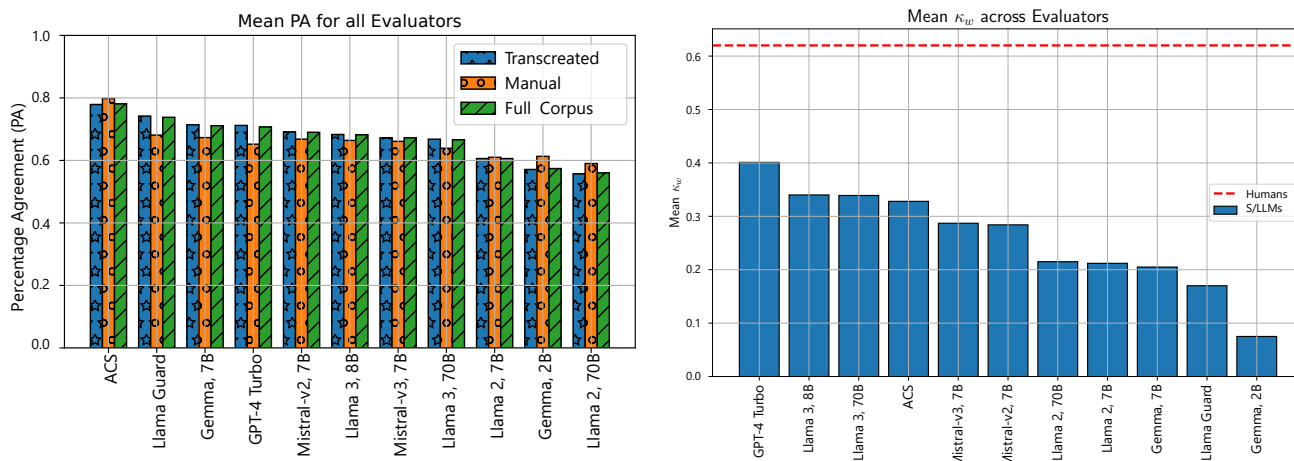


Figure 3: We labelled the prompt subset with the S/LLMs and compared their output with the majority vote of the human scores. In terms of raw accuracy (*left*), Llama Guard outperformed all other S/LLMs, closely followed by Gemma 7B and GPT-4 Turbo. ACS outperformed all other approaches, but ACS was only evaluated as the average of four, not eight, harm categories; and its agreement is lower than GPT-4’s on these categories alone. When looking at mean κ_w (*right*), it is clear that raw accuracy scoring is not a sufficient measure due to RTP-LX’s class imbalance—a lazy learner could output always the same label and obtain a decent performance. In fact, that is what happened for some models, such as Gemma 2B.

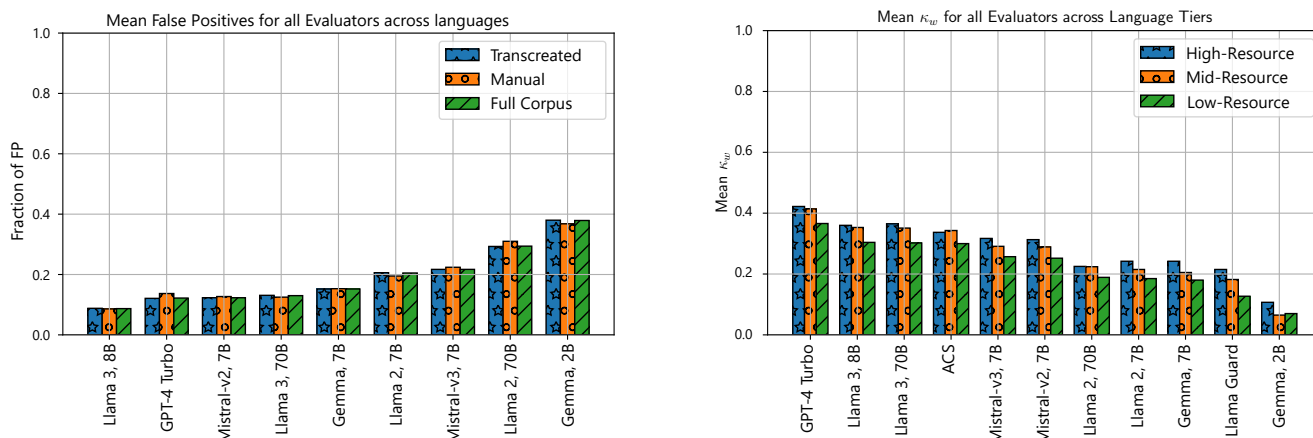


Figure 4: FPs across all languages for the S/LLMs. Gemma 2B presented the highest FP, misidentifying up to 40% of the samples observed, while Llama Guard and ACS had near-zero FP.

In terms of coverage, RTP-LX lacks sufficient dialectal coverage. Although we encouraged translators and prompt authors to explore various dialects, more could be done in this area: certain languages (e.g. Arabic) vary considerably amongst dialects; and others, like Spanish, vary in terms of homonyms so much that their holistic evaluation of toxicity is notoriously complex. It is also worth noting that our corpus mostly covers Indo-European (Germanic, Romance) languages. We plan to perform further additions to the corpus to explore other families and expand on dialectal differ-

Figure 5: Language availability versus κ_w over all languages in the prompts subset. All S/LLMs decreased in κ_w from high to low-resource languages, with differences of up to around 10%.

ences. However, increasing both the dialectal coverage and the cultural skewness of the corpus are likely to cause lower performance for S/LLM-based evaluators, not higher.

Toxic Language in a Global Context The study of toxicity in a multilingual setting is difficult not only due to the scarcity of quality, human-generated corpora especially designed for this task; but also because of the constant evolution of language and its perception by native speakers. It is also worth noting that things that may be offensive to a native speaker in one geo-locale may not be offensive to that

in another. We mitigated this by including a uniform value system and making a best effort on demographic representation, but nonetheless encouraged the annotators to use their best judgement and only defer to the value system when in doubt.

S/LLMs Evaluated S/LLMs are known to have data contamination issues that hamper fair evaluations. Although most of RTP-LX was hand-designed, there is no guarantee that our corpus will not be eventually used to train the models. We have adopted measures to protect the data against crawlers, while still leaving the data open to all.

Moreover, S/LLMs undergo frequent updates. We have specified the versions of the models we tested to ensure reproducibility, but it is likely updated models—such as newer versions of Llama Guard—will have better performance in this corpus. Likewise, we did not evaluate fine-tuned models. Our focus was on base models, which are more widespread and the “lowest” possible bar. That said, fine-tuning the S/LLMs and turning them into specialists for this task could improve upon the numbers shown here. The difficulty of fine-tuning lies on the scarcity of high-quality, available data: a common problem in the study of toxic language in NLP (Hartvigsen et al. 2022). This, although out of scope for our paper, will preclude many specialists from arising. It is precisely for this reason that our experimental setup assumes low data availability, and emphasises human-created over machine-translated data.

8 Conclusion

In this paper we explored the question of whether we can scale safety as fast as we scale S/LLMs. We framed it as a S/LLM capability analysis, by evaluating whether these models could work as annotators for multilingual toxic-language detection. To this end we introduced a human-annotated and human-transcreated corpus designed specifically to capture toxicity in a multilingual scenario.

We evaluated ten S/LLMs in RTP-LX, and found that they are able to score highly when measuring raw accuracy. However, by changing the metric (by comparing the outputs to human annotator judgements) showed that this performance did not necessarily mean that the models are reliable judges of toxicity. Indeed, we attributed that high accuracy to class imbalance, since the vast majority of the corpus is harmful. We noted that the agreement between S/LLMs and human annotators was relatively low, and often came coupled with an increasing tend to over-block benign content. This indicates a potentially poor understanding of the task by the S/LLM. One possible exception was GPT-4 Turbo, which was able to score within one standard deviation of human judgements.

Additionally, we found two pathologies common to some, if not all, S/LLMs evaluated: a tendency to select high-valued labels, which in RTP-LX means “extreme harm”, and low agreement with humans in context-dependent, subtle-yet-harmful content (e.g. microaggressions, bias). Both pathologies imply that the deployment of base, un-finetuned S/LLMs as multilingual harm detectors are likely to cause further problems, such as erasure.

Further work will scale RTP-LX to more dialects and language families. As mentioned earlier, transcreating and annotating the dataset with humans is comparatively time-consuming. However, it allows us to reliably address cultural sensitivity and subtler issues such as microaggressions. This means it is a worthwhile investment—as shown here, in a multilingual context, these subtle issues are precisely the ones requiring more attention in the field. Since we believe that RTP-LX is a resource necessary to combat harmful and toxic content in S/LLMs in research and in production, we have released the corpus, prompts, code used for this paper, and the rubrics, to the wider community.

A Annotation Details

Human involvement of RTP-LX involved two steps: transcreation, and annotation. The transcreation was performed by a single professional translator per-language, and they were paid at a varying rate depending on locale and seniority, between 19-54 USD/hr. The annotation was performed by three professional annotators, based on the criteria shown in Table 2. They were paid at a varying rate, likewise between 10-46.5 USD/hr. None of the annotators were transcreators. Prior to starting the work, we requested that the translators and annotators ensured that they took breaks and prioritized their own well-being. We performed several guideline fine-tuning rounds and worked closely with the annotation services company to ensure all doubts were cleared prior to starting the work. The final guidelines for both steps are in the repository.

A best effort was made to ensure representation in the annotation pool. From the optional exit survey, 52% themselves as female (r. 35% male); 6% identified as LGBTQ+; and 12% identified as neurodivergent or maybe neurodivergent. In terms of leaning for politics, 54% of the respondents identified themselves as centre-leaning (versus 22% left; 5% right). 20% of the respondents considered religion to be between important to extremely important to them, while 34% answered not important at all (32% somewhat important). Given that the English subset of RTP-LX dealt with sports in addition to religion and politics, we asked how important was sports to them. 39% of the respondents ranked sports between important to extremely important, while 19% said that it was not important at all and 28% mentioned it as somewhat important. The response rate for the survey was 88%, and all results from the survey, transcreation, and annotation work were anonymised. No particular individual can be identified based on the content.

B Languages Supported

The languages supported by RTP-LX, at the time of writing the first version of this paper, are in Table 1. The distinction between a dialect and a language may be a sensitive subject, and sometimes definition (consider Arabic, for example). In RTP-LX, sometimes we do not specify a dialect, but the corpus covers subjects relevant to the locations where it is spoken (for example, Kiswahili in Tanzania and in Kenya, Russian in Russia and Ukraine). For simplicity, we centralised all languages and dialects available in RTP-

Language	Primary Family	Subdivision	Dialects explicitly marked in RTP-LX	Language Class
Arabic	Afro-Asiatic	Central Semitic	Egyptian, Levantine, Saudi	5
Hebrew	Afro-Asiatic	Northwest Semitic	None	3
Indonesian	Austronesian	Malayo-Polynesian	None	3
Danish	Indo-European	North Germanic	None	3
Norwegian	Indo-European	North Germanic	None, but the corpus contains only Bokmål	1
Swedish	Indo-European	North Germanic	None	4
Dutch	Indo-European	West Germanic	None	4
English	Indo-European	West Germanic	None	5
German	Indo-European	West Germanic	None	5
Russian	Indo-European	East Slavic	None	4
Ukrainian	Indo-European	East Slavic	None	3
Czech	Indo-European	West Slavic	None	4
Polish	Indo-European	West Slavic	None	4
BCMS	Indo-European	South Slavic	Serbian	4
Spanish	Indo-European	Romance	None, focused on Peninsular Spanish	5
Portuguese	Indo-European	Romance	Brazilian and European	4
French	Indo-European	Romance	None, focused on Metropolitan French	5
Italian	Indo-European	Romance	None	4
Hindi	Indo-European	Hindustani	None	4
Thai	Kra-Dai	Tai	None	3
Kiswahili	Niger-Congo	Bantu	None	2
Standard Chinese	Sino-Tibetan	Sinitic	None (simplified and traditional scripts)	1
Japanese	Japonic	-	None	5
Korean	Koreanic	-	None	4
Turkish	Turkic	-	None	4
Finnish	Uralic	Finnic	None	4
Hungarian	Uralic	Ugric	None	4

Table 1: Languages supported by RTP-LX at the time of writing this paper. Sometimes a dialect is not specified, but the corpus covers subjects relevant to multiple locations where it is spoken (e.g., Russian in Russia and Ukraine). For simplicity, we centralised languages and the dialects available in RTP-LX to their main language (e.g., BCMS), with the exception of distinct writing systems (e.g., simplified vs traditional standard Chinese). The language class is a natural number between 0 and 5 denoting its level of online availability, with 5 being highest.

Criterion	Description	Scale
Toxicity	A holistic view of the content. How toxic would you rate the content to be?	1-5
Identity Attack	Hateful or hurtful comments targeting a group or person belonging to a group, ranging from stereotyping, denigration, and stigmatisation to explicitly harmful discourse (hate speech, slurs, etcetera).	1-3
Bias	Content containing or amplifying bias towards or against a specific group, ideology, or person. This is not necessarily toxic language.	1-3
Microaggression	Positive stereotyping and subtle yet not explicitly harmful content targeting a specific group (“where are you really from?”, “lots of Asians in my class, the curve will be dragged up” e.g.)	1-3
Violence	Content glorifying or promoting violence, threats, or self-harm.	1-3
Self-Harm	Content calling for actions intended to hurt, injure, or kill oneself.	1-3
Insult	Insulting or otherwise negative content, including profanities.	1-3
Sexual Content	Referring to sexual acts and related content, both consensual and not.	1-3

Table 2: Categories with definitions and scales for RTP-LX used in the annotation and in the prompts for the models.

LX to their main language. This meant that we used, for example, the linguistic classification for Serbian, Bosnian, Croatian, and Montenegrin (BCMS) as opposed to treating them separately. The only exception to this rule was if the writing system was noticeably different, such as in simplified versus traditional standard Chinese. The rationale here is that S/LLMs are primarily lexical engines.

C Select Results

In this section we show include selected results from our evaluations. We refer the reader to the repository for the full results and code to reproduce the work.

In terms of IRR (6), GPT-4 Turbo and Llama 3 8B were closest to the annotator’s judgements, off by an average Cohen’s κ_w difference of 0.22 and 0.28 points, respectively.

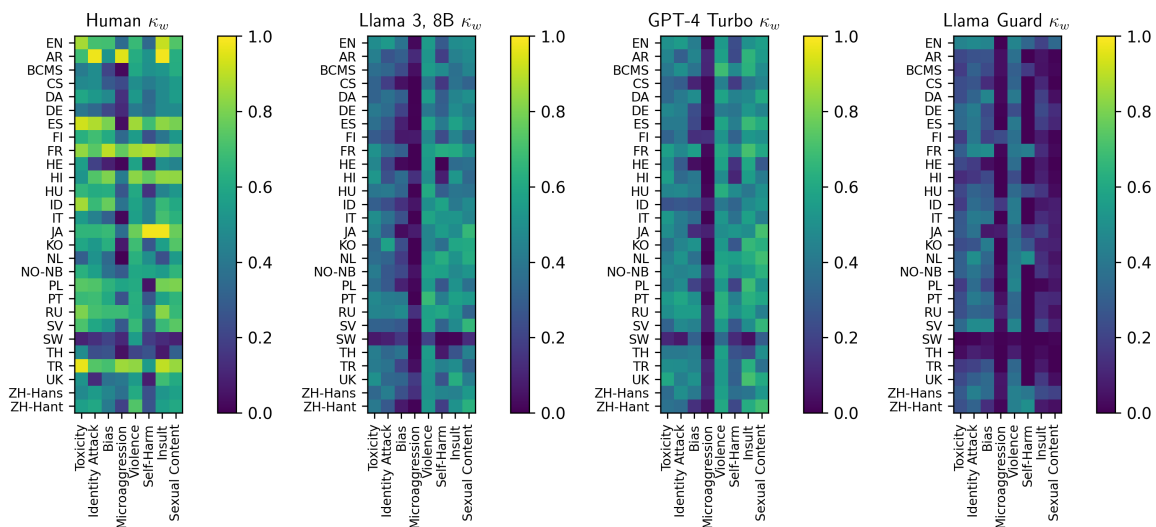


Figure 6: Cohen’s κ_w for RTP-LX for human annotators, GPT-4 Turbo, Llama 2 70B, and Llama Guard. IRR for human annotators is computed between the annotators. IRR for the S/LLMs is computed against the human majority vote. GPT-4 Turbo had the closest average IRR to human annotators (0.22 points difference). It noticeably outperformed other models some categories, but, like all models, it failed at detecting microaggressions. Llama Guard, which had topped PA scores, dropped to the second-to-last in terms of average κ_w . This suggests that it is a lazy learner, and it could overblock across all metrics.

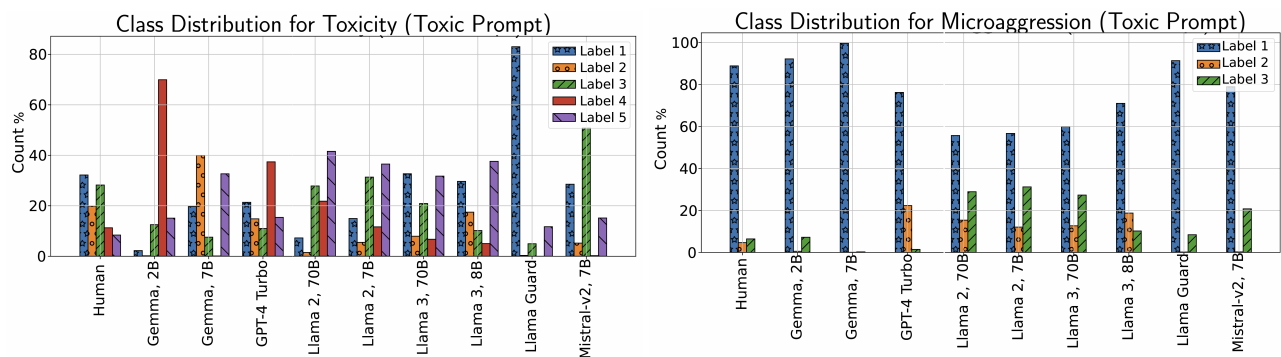


Figure 7: Class distribution comparing Toxicity and Microaggression for all models. The S/LLMs often output higher-valued labels, as opposed to positive labels. This was especially visible in Toxicity and Identity Attack. Likewise, the models often picked binary labels (e.g., Bias), although some models (Mixtral, Llama Guard) presented this pathology more than others.

For reference, the average IRR for human annotators was 0.62 ± 0.2 across all datasets. The Mistral models had closer IRRs to the human average (absolute differences of 0.34 and 0.33, respectively) than the Gemma models (0.54 for 2B and 0.41 for 7B). Gemma 2B was the lowest-performing model we tested. Both Llama-3 models tested were more adept at recognizing toxicity, identity attacks, and bias than their older versions, but failed at detecting self-harm. None of the Llama-2 and 3 models were able to detect microaggressions well. The differences in agreements with respect to human annotators were 0.28, 0.28, 0.41 and 0.40 for Llama-2 7B, *-2 70B, *-3 8B, and *-3 70B, respectively. Noticeably, increased model size did not enhance performance, but changes of model version did.

A breakdown by label output per-prompt (Figure 7) showed that the models preferred to output a higher-valued

label, yielding a high amount of FPs on the benign subset. This was evident in the categories where the models underperformed the most (Microaggression, Bias, Identity Attack), particularly for Gemma 2B and Mistral v3, which deemed benign completions to be extremely toxic. In Self-Harm and Sexual Content, all models did well, with ACS overblocking slightly more than other models. In Violence and other categories, most models output a binary label, which meant that the S/LLMs were less prone to consider context when emitting a judgement. Toxicity especially caused a disparity in some models (e.g., Gemma and Mixtral), where larger labels were often observed, while Llama Guard was unable to detect it, same as Insult.

Ethics Statement

All aspects of this research were reviewed and approved by the Institutional Review Board of our organisation. Given the nature of the corpus and the tendency for S/LLMs to memorise highly available data, reproducibility—a cornerstone of science—becomes complex. To ensure RTP-LX remains a resource for the community without allowing S/LLMs to propagate this content, we have taken measures to protect the data by password-protecting it.

Acknowledgments

We would like to thank Can Gören for his contribution to the Turkish subset of RTP-LX; and the anonymous reviewers, whose comments drastically improved this paper.

References

- Ahuja, S.; Aggarwal, D.; Gumma, V.; Watts, I.; Sathé, A.; Ochieng, M.; Hada, R.; Jain, P.; Ahmed, M.; Bali, K.; and Sitaram, S. 2024. MEGEVERSE: Benchmarking Large Language Models Across Languages, Modalities, Models and Tasks. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proc Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2598–2637. Mexico City, Mexico: Association for Computational Linguistics.
- AI@Meta. 2024. Llama 3 Model Card.
- Carlini, N.; Ippolito, D.; Jagielski, M.; Lee, K.; Tramer, F.; and Zhang, C. 2023. Quantifying Memorization Across Neural Language Models. In *Int Conf on Learning Representations*.
- Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; Ye, W.; Zhang, Y.; Chang, Y.; Yu, P. S.; Yang, Q.; and Xie, X. 2024. A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.*
- Costa-jussà, M. R.; Cross, J.; Çelebi, O.; Elbayad, M.; Heafield, K.; Heffernan, K.; Kalbassi, E.; Lam, J.; Licht, D.; Maillard, J.; Sun, A.; Wang, S.; Wenzek, G.; Youngblood, A.; Akula, B.; Barrault, L.; Gonzalez, G. M.; Hansanti, P.; Hoffman, J.; Jarrett, S.; Sadagopan, K. R.; Rowe, D.; Spruit, S.; Tran, C.; Andrews, P.; Ayan, N. F.; Bhosale, S.; Edunov, S.; Fan, A.; Gao, C.; Goswami, V.; Guzmán, F.; Koehn, P.; Mourachko, A.; Ropers, C.; Saleem, S.; Schwenk, H.; and Wang, J. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation.
- Davidson, T.; Bhattacharya, D.; and Weber, I. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In Roberts, S. T.; Tetreault, J.; Prabhakaran, V.; and Waseem, Z., eds., *Proc of the Third Workshop on Abusive Language Online*, 25–35. Florence, Italy: Association for Computational Linguistics.
- De Wuyter, A.; Wang, X.; Sokolov, A.; Gu, Q.; and Chen, S.-Q. 2023. An evaluation on large language model outputs: Discourse and memorization. *Natural Language Processing Journal*, 4: 100024.
- Deng, Y.; Zhang, W.; Pan, S. J.; and Bing, L. 2024. Multilingual Jailbreak Challenges in Large Language Models. arXiv:2310.06474.
- Dhamala, J.; Sun, T.; Kumar, V.; Krishna, S.; Pruksachatkun, Y.; Chang, K.-W.; and Gupta, R. 2021. BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. In *Proc. of the 2021 ACM Conf on Fairness, Accountability, and Transparency, FAccT '21*, 862–872. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.
- Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; and Smith, N. A. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3356–3369. Online: Association for Computational Linguistics.
- Hada, R.; Gumma, V.; de Wuyter, A.; Diddee, H.; Ahmed, M.; Choudhury, M.; Bali, K.; and Sitaram, S. 2024. Are Large Language Model-based Evaluators the Solution to Scaling Up Multilingual Evaluation? In Graham, Y.; and Purver, M., eds., *Findings of the Association for Computational Linguistics: EACL 2024*, 1051–1070. St. Julian's, Malta: Association for Computational Linguistics.
- Hamad, N.; Jarrar, M.; Khalilia, M.; and Nashif, N. 2023. Offensive Hebrew Corpus and Detection using BERT. In *The 20th ACS/IEEE Int Conf on Computer Systems and Applications*. Egypt: IEEE.
- Hartvigsen, T.; Gabriel, S.; Palangi, H.; Sap, M.; Ray, D.; and Kamar, E. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proc of the 60th Annual Meeting of the Association for Computational Linguistics*, 3309–3326. Dublin, Ireland: Association for Computational Linguistics.
- Holtermann, C.; Röttger, P.; Dill, T.; and Lauscher, A. 2024. Evaluating the Elementary Multilingual Capabilities of Large Language Models with MultiQ. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 4476–4494. Bangkok, Thailand: Association for Computational Linguistics.
- Inan, H.; Upasani, K.; Chi, J.; Rungta, R.; Iyer, K.; Mao, Y.; Tontchev, M.; Hu, Q.; Fuller, B.; Testuggine, D.; and Khabsa, M. 2023. Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations. ArXiv:2312.06674.
- Jain, D.; Kumar, P.; Gehman, S.; Zhou, X.; Hartvigsen, T.; and Sap, M. 2024. PolygloToxicityPrompts: Multilingual Evaluation of Neural Toxic Degeneration in Large Language Models. In *First Conference on Language Modeling*.
- Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. 2310.06825:arXiv.
- Joshi, P.; Santy, S.; Budhiraja, A.; Bali, K.; and Choudhury, M. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proc of the 58th Annual Meeting of the Association for Computational Linguistics*.
- Lai, V.; Ngo, N.; Pourn Ben Veyseh, A.; Man, H.; Derroncourt, F.; Bui, T.; and Nguyen, T. 2023a. ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 13171–13189. Singapore: Association for Computational Linguistics.
- Lai, V. D.; Ngo, N. T.; Veyseh, A. P. B.; Man, H.; Derroncourt, F.; Bui, T.; and Nguyen, T. H. 2023b. ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Lee, J.; Le, T.; Chen, J.; and Lee, D. 2023. Do Language Models Plagiarize? In *Proceedings of the ACM Web Conference 2023, WWW '23*, 3637–3647. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394161.

- Leite, J. A.; Silva, D. F.; Bontcheva, K.; and Scarton, C. 2020. Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis. In *AAACL-IJCNLP*.
- Li, J.; Wang, J.; Hu, J.; and Jiang, M. 2024. How Well Do LLMs Identify Cultural Unity in Diversity? In *First Conference on Language Modeling*.
- Moon, J.; Cho, W. I.; and Lee, J. 2020. BEEP! Korean Corpus of Online News Comments for Toxic Speech Detection. In *Proc of the Eighth Int Workshop on Natural Language Processing for Social Media*, 25–31. Online: Association for Computational Linguistics.
- OpenAI. 2024. GPT-4 Technical Report. arXiv:2403.08295.
- Rauh, M.; Mellor, J.; Uesato, J.; Huang, P.-S.; Welbl, J.; Weidinger, L.; Dathathri, S.; Glaese, A.; Irving, G.; Gabriel, I.; Isaac, W.; and Hendricks, L. A. 2022. Characteristics of harmful text: Towards rigorous benchmarking of language models.
- Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; and Smith, N. A. 2019. The Risk of Racial Bias in Hate Speech Detection. In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proc. of the 57th Annual Meeting of ACL*, 1668–1678. Florence, Italy: Association for Computational Linguistics.
- Sheng, E.; Chang, K.-W.; Natarajan, P.; and Peng, N. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proc of the 2019 Conf on Empirical Methods in Natural Language Processing and the 9th Int Joint Conf on Natural Language Processing*, 3407–3412. Hong Kong, China: Association for Computational Linguistics.
- Sigurbjergsson, G. I.; and Derczynski, L. 2020. Offensive Language and Hate Speech Detection for Danish. In *Proc of the Twelfth Language Resources and Evaluation Conference*, 3498–3508. Marseille, France: European Language Resources Association. ISBN 979-10-95546-34-4.
- Team, G.; Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivièrè, M.; Kale, M. S.; Love, J.; Tafti, P.; Hussenot, L.; Sessa, P. G.; Chowdhery, A.; Roberts, A.; Barua, A.; Botev, A.; Castro-Ros, A.; Slone, A.; Héliou, A.; Tacchetti, A.; Bulanov, A.; Paterson, A.; Tsai, B.; Shahriari, B.; Lan, C. L.; Choquette-Choo, C. A.; Crepy, C.; Cer, D.; Ippolito, D.; Reid, D.; Buchatskaya, E.; Ni, E.; Noland, E.; Yan, G.; Tucker, G.; Muraru, G.-C.; Rozhdestvenskiy, G.; Michalewski, H.; Tenney, I.; Grishchenko, I.; Austin, J.; Keeling, J.; Labanowski, J.; Lepiau, J.-B.; Stanway, J.; Brennan, J.; Chen, J.; Ferret, J.; Chiu, J.; Mao-Jones, J.; Lee, K.; Yu, K.; Millican, K.; Sjoesund, L. L.; Lee, L.; Dixon, L.; Reid, M.; Miłkuła, M.; Wirth, M.; Sharman, M.; Chirnaev, N.; Thain, N.; Bachem, O.; Chang, O.; Wahltinez, O.; Bailey, P.; Michel, P.; Yotov, P.; Chaabouni, R.; Comanescu, R.; Jana, R.; Anil, R.; McIlroy, R.; Liu, R.; Mullins, R.; Smith, S. L.; Borgeaud, S.; Girgin, S.; Douglas, S.; Pandya, S.; Shakeri, S.; De, S.; Klimenko, T.; Hennigan, T.; Feinberg, V.; Stokowiec, W.; hui Chen, Y.; Ahmed, Z.; Gong, Z.; Warkentin, T.; Peran, L.; Giang, M.; Farabet, C.; Vinyals, O.; Dean, J.; Kavukcuoglu, K.; Hassabis, D.; Ghahramani, Z.; and et al., D. E. 2024. Gemma: Open Models Based on Gemini Research and Technology. arXiv:2403.08295.
- Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *ArXiv*, abs/2307.09288.
- Wang, B.; Chen, W.; Pei, H.; Xie, C.; Kang, M.; Zhang, C.; Xu, C.; Xiong, Z.; Dutta, R.; Schaeffer, R.; Truong, S. T.; Arora, S.; Mazeika, M.; Hendrycks, D.; Lin, Z.; Cheng, Y.; Koyejo, S.; Song, D.; and Li, B. 2023. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. In *NeurIPS 2023*.
- Wang, W.; Tu, Z.; Chen, C.; Yuan, Y.; Huang, J.-t.; Jiao, W.; and Lyu, M. 2024a. All Languages Matter: On the Multilingual Safety of LLMs. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 5865–5877. Bangkok, Thailand: Association for Computational Linguistics.
- Wang, Y.; Zhai, C.; and Hassan, H. 2020. Multi-task Learning for Multilingual Neural Machine Translation. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proc of the 2020 Conf on Empirical Methods in Natural Language Processing*, 1022–1034. Online: Association for Computational Linguistics.
- Wang, Y.; Zhai, Z.; Li, H.; Han, X.; Lin, S.; Zhang, Z.; Zhao, A.; Nakov, P.; and Baldwin, T. 2024b. A Chinese Dataset for Evaluating the Safeguards in Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 3106–3119. Bangkok, Thailand: Association for Computational Linguistics.
- Wei, F.; Chen, X.; and Luo, L. 2024. Rethinking Generative Large Language Model Evaluation for Semantic Comprehension. *ArXiv*, abs/2403.07872.
- Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In Oh, A.; Neumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *NeurIPS*, volume 36, 46595–46623. Curran Associates, Inc.