The State of Multilingual LLM Safety Research: From Measuring The Language Gap To Mitigating It

Zheng-Xin Yong[1], **Beyza Ermis**[2], **Marzieh Fadaee**[2],
Stephen H. Bach[1], **Julia Kreutzer**[2]

[1]Brown University, [2]Cohere Labs

Correspondence: contact.yong@brown.edu, juliakreutzer@cohere.com

Abstract

This paper presents a comprehensive analysis of the linguistic diversity of LLM safety research, highlighting the English-centric nature of the field. Through a systematic review of nearly 300 publications from 2020–2024 across major NLP conferences and workshops at *ACL, we identify a significant and growing language gap in LLM safety research, with even high-resource non-English languages receiving minimal attention. We further observe that non-English languages are rarely studied as a standalone language and that English safety research exhibits poor language documentation practice. To motivate future research into multilingual safety, we make several recommendations based on our survey, and we then pose three concrete future directions on safety evaluation, training data generation, and crosslingual safety generalization. Based on our survey and proposed directions, the field can develop more robust, inclusive AI safety practices for diverse global populations.

Content Warning: This paper contains examples of harmful language.

1 Introduction

The rapid advancement of large language models (LLMs) has transformed the artificial intelligence landscape, enabling increasingly sophisticated capabilities across domains including healthcare (Singhal et al., 2023; Nazi and Peng, 2024; Singhal et al., 2025), education (Neumann et al., 2024; Zhang et al., 2024b; Wen et al., 2024), and media content generation (Wang et al., 2023; Zhang et al., 2024a; Barman et al., 2024). As these powerful systems are deployed globally and used across different linguistic communities (Tamkin et al., 2024), ensuring their safe and secure operation across diverse linguistic and cultural contexts has emerged as a critical research imperative. While significant progress has been made in developing safety mechanisms for high-resource languages (Shi et al., 2024a; Dong et al., 2024), particularly English, the multilingual dimensions of LLM safety remain considerably underexplored. For example, *all* the public safety evaluation datasets reviewed by Dong et al. (2024) include English content, with only two datasets being bilingual (English and Chinese). This gap creates potentially dangerous blind spots in our safety frameworks and raises fundamental questions about the equitable distribution of AI benefits and risks (Yong et al., 2023a; Ermis et al., 2024; Aakanksha et al., 2024; Kanepajs et al., 2024; Bengio et al., 2025).

Multilingual LLM safety encompasses challenges that extend well beyond the simple translation of existing safety techniques. Languages differ not only in their vocabulary and grammatical structures but also in their cultural connotations (Hoijer, 1954; Jiang, 2000; Everett, 2012; Kramsch, 2014; Mazari and Derraz, 2015), metaphorical expressions (Saygin, 2001; Khoshtab et al., 2025), taboos (Dewaele, 2004), and social norms (Sridhar, 1996; Baquedano-López and Kattan, 2007; Fasya and Sari, 2021). Therefore, content that is harmless in one cultural context may be deeply offensive or harmful in another (Keipi et al., 2016; Ermis et al., 2024; Aakanksha et al., 2024; Korre et al., 2025), or vice versa. For instance, in South-East Asia, the term "*banana*"—which connotes "yellow on the outside, white on the inside"—is used to disparage people of Asian descent who are perceived as forgoing their cultural identity and having adopted Western cultural values and behaviors (Khoo, 2003; Trieu, 2019). On the other hand, the Chinese word 屌, which literally translates as "dick", can be used in both offensive (i.e., swear words) and non-offensive (i.e., an adjective to praise someone who possesses a remarkable talent) settings (Carson and Jiang, 2021).

The wide disparity in language resources (Joshi et al., 2020; Nigatu et al., 2024)—from high-resource languages like English, Mandarin,

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 15845–15860
November 4-9, 2025 ©2025 Association for Computational Linguistics

Categories	Definitions	Examples
Jailbreaking attacks	Work on designing adversarial prompts to bypass refusal safety guardrails or detecting jailbreaking attacks	Zeng et al. (2024), Wang et al. (2024c)
Toxicity and bias	Work on toxic content and stereotypical bias in training data and output generations	Zhu et al. (2024), Kim et al. (2024)
Factuality and hallucination	Work on nonsensical, unfaithful, and factually incorrect content generated by LLMs	Pal and Sankarasubbu (2024)
AI privacy	Work on memorization, private data leakage, and unlearning	Dou et al. (2024), Shi et al. (2024b)
Policy	Work on governance frameworks, regulatory approaches, and ethical guidelines for responsible AI deployment	Goanta et al. (2023)
LLM alignment	Work that spans multiple subtopics above or is related to other LLM safety subtopics such as RLHF alignment algorithms	Wang et al. (2024d), Yang et al. (2024b)
Not related to safety	Work that does not belong to any of the topics above	Manino et al. (2022)

Table 1: Categorization of papers in our LLM safety survey study.

and Spanish to thousands of low-resource languages—creates uneven safety landscapes with potentially severe consequences for marginalized linguistic communities. Several commercial LLMs have demonstrated significantly weaker safety performance when prompted in non-English languages, producing harmful content and undesirable outputs that would be filtered in English contexts (Yong et al., 2023a; Deng et al., 2024; Wang et al., 2024a; Al Ghanim et al., 2024; Yoo et al., 2024; He et al., 2024; Shen et al., 2024; Nigatu and Raji, 2024; Poppi et al., 2025; Aakanksha et al., 2024; Jain et al., 2024; Chan et al., 2025). These disparities in safety protections, combined with increasingly capable LLMs, risk magnifying societal harms within multilingual communities. While companies behind frontier LLMs have taken concerted efforts to perform multilingual safety alignment training and red-teaming (Grattafiori et al., 2024; Cohere et al., 2025; OpenAI, 2025), these initiatives remain limited in scope. For instance, among the top-ranking LLMs on Chatbot Arena—a widely used leaderboard platform for evaluating LLMs through user-submitted preference—20 of 24 of those that provide a system report have wide multilingual support, but *only 5* reported multilingual safety alignment training and red-teaming efforts. This *gap between multilingual deployment capabilities and safety alignment* calls for further participation from both private enterprises and academia on multilingual safety alignment.

We perform a systematic review of nearly 300 LLM safety publications over the past five years in ACL proceedings (Section 2), and we uncover a concerning trend: the vast majority of safety research is centered on English-language models, while comparatively little work addresses safety in non-English or multilingual contexts. This imbalance has become more pronounced over time. Even Mandarin Chinese—the second most studied language—still has about *ten times less* research than English. This disparity persists across multiple subdomains of safety research. Furthermore, non-English languages are rarely studied as a standalone language but rather as part of broader multilingual evaluations, which often lack the nuance and depth necessary to address language-specific safety challenges and cultural contexts. Lastly, we discover that only half of English safety research publications document the limitations of their language coverage.

These findings highlight critical gaps in the current landscape of LLM safety research and motivate the need for more targeted efforts to address multilingual safety concerns. To help close this gap, we outline three tractable directions for future multilingual safety work: (1) developing culturally grounded evaluation benchmarks, (2) curating diverse multilingual safety training data, and (3) deepening our understanding of alignment challenges across languages.

2 The Language Gap in LLM Safety Research

To understand the language gap in LLM safety research, we systematically survey relevant papers and analyze how safety research is distributed across languages and subtopics, as well as how non-English language research is conducted and reported.

2.1 Methodology

We collect work related to LLM safety and manually annotate the languages studied in each paper,

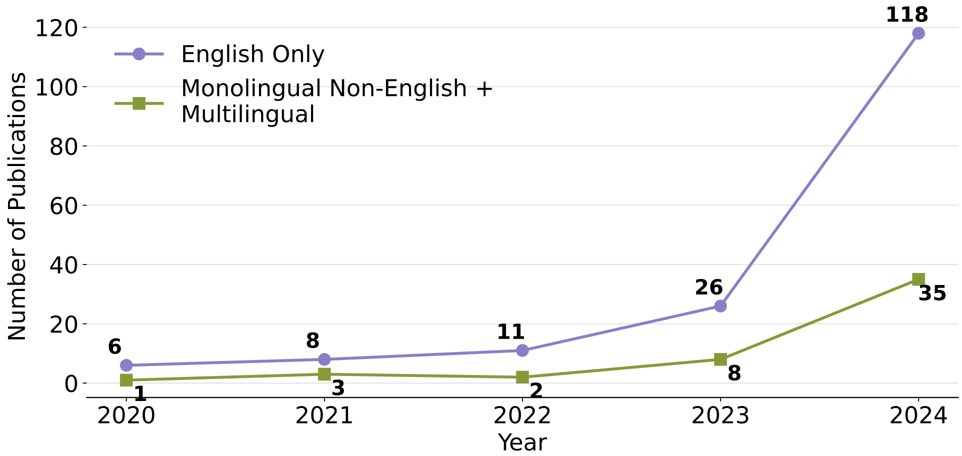

Figure 1: Trends of English-only and multilingual LLM safety publications in *ACL conferences and workshops over the past five years: the language gap in LLM safety research widens.

along with their safety subtopic. To reduce human annotation efforts while ensuring that our findings reflect the overall trends in the field, we perform the following strategies:

1. **Venue selection**: We focus on all *ACL venues such as ACL and EMNLP, including both conferences and workshops, as we believe they are the venues with the most linguistically diverse NLP works compared to other venues such as ICLR, NeurIPS, and ICML.

2. **Keyword filter**: We filter the safety-related publications through keyword matching with words "safe" and "safety" in paper abstracts. Using these two terms we get a good proxy for the distribution of diverse LLM safety literature.

3. **Manual categorization**: We adopt a simplified categorization following Cui et al. (2024), which is representative of the type of safety work published at *ACL, and we manually categorize publications into seven different subtopics as shown in Table 1.

4. **Language Documentation**: We annotate the languages that each work addresses[1], and we indicate if the language(s) studied are mentioned in the work. We group them into three categories: monolingual English, monolingual non-English, and multilingual (covering two or more languages).

Annotation Task	**Type**	**Avg**	**Std**
Safety topic	Categorical	0.83	0.19
Has non-English?	Binary	0.81	0.15
Specifies languages?	Binary	0.80	0.04
Covered languages	List	0.96	0.05

Table 2: Average and standard deviation of four pairwise annotator comparisons. Agreement on 'language coverage' is measured with Jaccard similarity, and all other categories are measured with Cohen's κ.

Annotations were manually performed by the authors. In total, we annotated nearly 300 publications from year 2020 till year 2024. Of these, 28% were false positives from our keyword matching process (i.e., unrelated to LLM safety), and were filtered out before we perform further analysis.

Table 2 reports the mean and standard deviation of pairwise inter-annotator agreement scores on subsets of 20 repeated annotations. We perform a 4 × 20 pairwise agreement study across distinct subsets to maximize the representativeness of our survey corpus and ensure robust assessment of annotation consistency. We find that inter-annotator agreement is consistently high, between 0.80 and 0.96 on average per category, but we note that the annotations may still contain imperfections.

2.2 Findings

English-centricity of LLM safety research. Figure 1 highlights a stark language imbalance in LLM safety research published at *ACL conferences and workshops over the past five years. The data reveals a clear English-centric pattern that has persisted

[1] If the languages studied were not explicitly mentioned, we followed up on their training and evaluation datasets to identify the language coverage of the work.

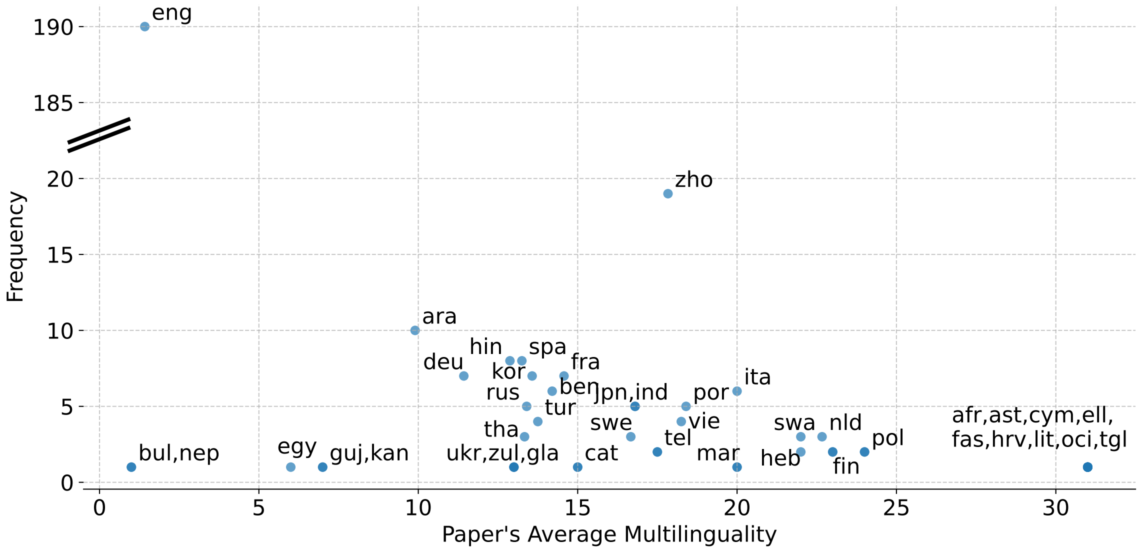

Figure 2: Measure of how often a language is studied ("Frequency") and the average number of languages covered by all papers in which the language appear in ("Paper's Average Multilinguality").

throughout this period. English-only research dominates across all years, with a particularly dramatic increase in recent publications. The trend shows consistent underrepresentation of multilingual non-English research, with the gap widening significantly over time. While both categories have grown as LLM safety has gained prominence, the proportional imbalance remains. English-only publications have consistently outnumbered multilingual and non-English work, and this absolute gap has widened over time, from 5 in 2020 to 83 in 2024. While both categories have grown, the increase is disproportionately concentrated in English-only research.

Non-English languages are studied in herds. Another aspect of the marginalization of non-English languages is that they are often addressed as part of large multilingual evaluations, rather than studied in depth on their own. In many cases, breadth is prioritized over depth, and multilingual studies are preferred over focused analyses of monolingual ones.[2]

This is shown in Figure 2 which provides a detailed breakdown of how frequently a language is studied (y-axis) and how often it is studied alongside other languages (x-axis). English (eng) exhibits overwhelming dominance with a frequency nearly ten times higher than Chinese (zho)—the second most studied language. However, English is primarily studied in isolation, resulting in a low average multilinguality score. In contrast, languages with moderate representation like Chinese (zho), Arabic (ara) and Spanish (spa) appear primarily in multilingual studies, suggesting that deeper, language-specific safety analyses remain limited even for widely spoken languages. This trend is even more noticeable for under-resourced languages such as Swahili (swa) and Telugu (tel), and especially for languages at the extreme end of the multilingualism spectrum such as Afrikaans (afr), which appears only in a single paper that covers approximately 30 languages (Guerreiro et al., 2023). Such inclusion severely limits the possibility for language-specific safety analysis and gaining meaningful insights. We commend focused analysis on individual lower-resource languages such as Nakov et al. (2021) and Niraula et al. (2021), who specifically study disinformation and offensive language detection in Bulgarian and Nepali social media, respectively.

Disparities in subtopics of safety. Breaking down LLM safety publications by specific safety subtopics in Figure 3(a), we find that English-centricity persists across all domains, with English-only publications substantially outnumbering multilingual work in every category. LLM alignment and jailbreaking attacks demonstrate the most pronounced disparities, suggesting that these critical safety areas receive particularly limited cross-

[2] Since our study only captures published papers, we might be missing out on rejected works. There may be a reviewer preference for multilingual over monolingual non-English papers.

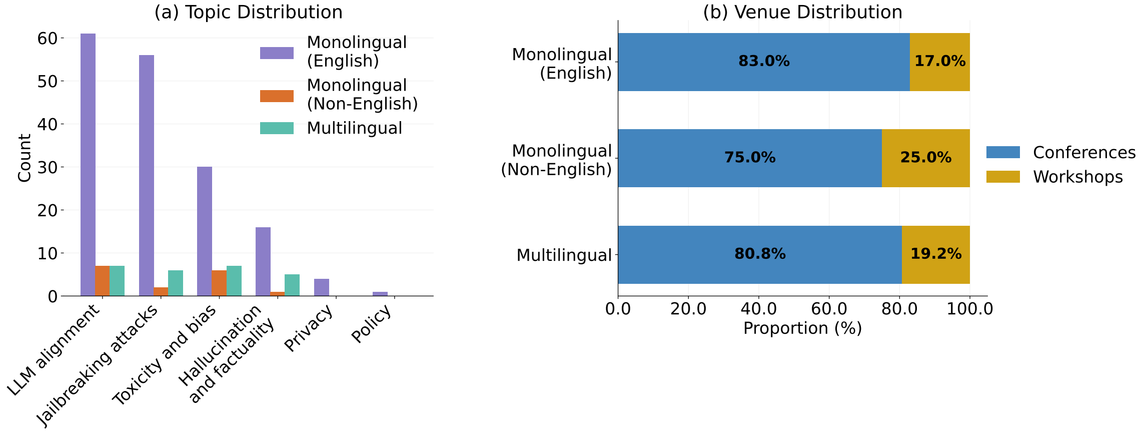

Figure 3: Distribution of LLM safety publications by (a) safety subtopics and (b) publication venues.

linguistic attention. In particular, LLM alignment work involving evaluation (Yuan et al., 2024; Hua et al., 2024; Hammoud et al., 2024; Gabriel et al., 2024) and algorithmic improvement (Zhou et al., 2024; Hassan et al., 2024) would benefit from further research with expanded language coverage. Toxicity and bias research shows a similar pattern despite being a domain where cultural and linguistic variations are especially relevant (Costa-jussà et al., 2023a; Tao et al., 2024; Devinney et al., 2024; Bhutani et al., 2024). The near absence of multilingual work in privacy and policy domains indicates these emerging safety concerns are being conceptualized almost exclusively through an English-language framework, potentially overlooking important cultural and legal variations that exist across different linguistic contexts (Larsen and Dignum, 2024).

Valuable role of workshops. Figure 3 (b) reveals an interesting pattern in the distribution of LLM safety publications across venue types. While conferences dominate across all language categories, monolingual non-English safety papers are 46% relatively more likely to appear in workshops than English-only papers, highlighting the valuable accessibility that workshops offer for this line of work. This suggests that workshops, such as Workshop on Gender Bias in Natural Language Processing (GeBNLP) and Workshop on Safety for Conversational AI (Safety4ConvAI), serve as more accessible venues for disseminating non-English safety research. Some possible reasons include non-English safety research may face a higher barrier to entry at prestigious conferences, or workshops are better for fostering emerging research areas. The pattern indicates that, beyond the overall English-centricity of safety research documented in previous figures, additional structural factors may be affecting how non-English safety work is evaluated and disseminated within the community.

Category	Does the paper mention languages studied?	
	No (↓)	Yes (↑)
Mono. English	50.6%	49.4%
Mono. Non-English	0.0%	100.0%
Multilingual	0.0%	100.0%

Table 3: Proportion of language documentation practice among LLM safety publications.

Language documentation practice differs for English-only research. We argue that it is important for LLM safety research to *explicitly document the languages studied* (also known as Bender's rule (Bender, 2011, 2019)) for two key reasons. (1) Safety alignment does not necessarily generalize across languages (Yong et al., 2023a; Wang et al., 2024b; Yoo et al., 2024; Al Ghanim et al., 2024). Clearly stating which languages were included enables future researchers to understand the specific linguistic contexts in which safety findings have been validated. (2) By explicitly acknowledging language limitations, the field can more accurately measure progress in expanding safety coverage across languages, thus encouraging a more equitable distribution of safety research to serve a broader range of global populations. Based on the data presented in Table 3, we observe substantially different patterns in language documentation practices across LLM safety publications. English-only publications show a concerning trend with 50.6% failing to explicitly name the language studied – in other words, "English" is not mentioned throughout the paper. In contrast, both non-English

Models	en	zh	fr	ru	de	ar	hi	es	ja	bn	Average ↑	Worst Case* ↑
ChatGPT (OpenAI, 2022)	99.0	91.9	86.3	87.5	85.3	90.8	81.7	91.5	79.0	62.6	**85.56**	62.6
PaLM-2 (Anil et al., 2023)	89.7	78.4	84.6	85.9	83.6	82.6	83.0	85.7	70.1	78.1	82.17	**70.1**
Llama-2 (Touvron et al., 2023)	85.4	73.5	83.2	82.3	82.0	-	63.5	79.3	71.0	-	77.53	63.5
Vicuna (Chiang et al., 2023)	94.0	89.4	90.6	83.3	88.3	43.4	36.8	88.8	60.2	18.4	69.32	**18.4 (!)**

Table 4: Harmlessness scores of different models across 10 languages, based on the results from (Wang et al., 2024a). We augment the original table with a new *"Worst Case"* column for the lowest harmlessness score. We use **bold text** to indicate the cases where average score is not necessarily aligned with worst-case score, and we use **red text and exclamation mark** to indicate how not reporting Worst-Case score can create a false sense of safety.

monolingual and multilingual publications demonstrate full compliance, with 100% explicitly documenting the languages studied. This disparity highlights a systematic bias in reporting practices, where English-centered research often proceeds under an implicit assumption of universality, whereas non-English research demonstrates greater methodological transparency.

2.3 Moving Forward for *ACL Venues

Our survey reveals that English safety research remains overwhelmingly dominant in nearly every dimension—publication volume, topical coverage, methodological reporting, and conference visibility. Nonetheless, Figure 1 shows an encouraging trend of growing multilingual safety research over time.

One concrete and low-effort step toward improving documentation is integrating language coverage reporting into *ACL proceedings. OpenReview submissions already include a metadata field where authors can indicate the languages studied, but this information is currently private. Making this metadata public would allow for more transparent tracking of linguistic representation and support future meta-analyses of multilingual research, particularly in the context of LLM safety.

Addressing the deeper structural imbalance in language and topic representation will require long-term efforts. We believe that conference and workshop organizers can provide incentive structures to address this systemic imbalance, such as special conference theme tracks dedicated to multilingual safety subtopics and/or creating shared workshop tasks on multilingual safety benchmarks. These initiatives could meaningfully expand the scope and visibility of research beyond English, helping the community better serve diverse user populations.

3 Future Research Directions for Multilingual LLM Safety

In addition to providing recommendations to *ACL organizers, we propose several key research priorities for researchers and model developers to advance multilingual LLM safety alignment.

3.1 Safety Evaluation for Multilingual Models

Moving beyond average safety criterion. Traditional evaluation metrics focus on average performance across languages, for which the model that maximizes the uniformly weighted average across tasks and languages is considered best. However, this criterion is susceptible to outliers (e.g., due to unsupported languages) and not suitable for comparing models with different language and task support (Kreutzer et al., 2025). In the context of multilingual safety, where reporting average scores is the norm (Guan et al., 2024), this matters even more since averaging might obscure critical safety failures.

To illustrate this blind spot, we add the additional *worst-case* harmlessness score metric to an ACL 2024 paper (Wang et al., 2024a) and report the results in Table 4. The table reveals two findings. First, if the winner were chosen based solely on the highest average harmlessness score, it would be ChatGPT (OpenAI, 2022), with a score of 85.56. However, its worst-case score (i.e., the lowest harmlessness score across languages) is only 62.6. This is notably lower than the worst-case score of PaLM-2 (70.1), despite PaLM-2 (Anil et al., 2023) having a lower average score (82.17). This discrepancy highlights that strong average performance does not necessarily reflect robustness in the worst-case scenarios. Second, and more importantly, despite a high average harmlessness score, Vicuna's (Chiang et al., 2023) worst-case harmlessness score is just 18.4 due to unsafe behaviour in Bengali (bn). This suggests that relying on average metrics alone may create a false sense of safety, potentially leading to the deployment of models like Vicuna in languages where they produce harmful content. In future work, we believe that, in addition to reporting worst-case performance to ensure that models meet fundamental safety thresholds across all languages, researchers should explore designing adaptive thresholding mechanisms that establish language-specific safety baselines according to their unique cultural contexts and user groups.

Wider language coverage in evaluation. We observe that current multilingual red-teaming practice mostly focuses on languages that models are finetuned on during post-pretraining processes, such as instruction-following and alignment finetuning (Üstün et al., 2024; Grattafiori et al., 2024). Given that language contamination in pretraining can facilitate crosslingual transfer (Blevins and Zettlemoyer, 2022), it raises valid concerns about whether *exempting* certain languages from the safety evaluation of multilingual LLMs is justified. Language exemptions risk creating blind spots in safety assessments precisely where they might be most needed, as models can bypass safety guardrails when prompted in languages underrepresented in pretraining (Shen et al., 2024). For instance, the Llama-3 model report presents red-teaming results for only eight languages (six of which are high-resource) (Grattafiori et al., 2024). Yet the strong multilingual model has been adapted for languages not covered in its safety evaluation, such as Indonesian (Huang et al., 2024b).

We urge researchers to develop more sophisticated evaluation protocols that can detect and account for potential contamination and to issue disclaimers when safety alignment has not been conducted in certain languages. This would help ensure that speakers of those languages are aware of potential risks. Such transparency would allow communities to make informed decisions about model deployment while encouraging greater accountability from developers to expand alignment efforts to underserved languages.

Incorporate diverse and natural linguistic patterns. We believe evaluating multilingual safety requires a fundamental shift away from treating evaluation as merely adding more languages to existing benchmarks, as they should incorporate linguistic patterns used by real-life speakers. One case study is *code-switching*—the communication pattern of alternating between languages within a single utterance (Nilep, 2006; Gardner-Chloros, 2009; Winata et al., 2023)—which is shown to be able to jailbreak multilingual safety guardrails (Yoo et al., 2024; Yang et al., 2024a; Song et al., 2025). Another example is Al Ghanim et al.'s (2024) discovery that while LLMs remain safe in standardized Arabic scripts, they are jailbroken when Arabic inputs are written in Arabizi form—a system of writing Arabic using English characters and commonly used among native speakers communicating digitally (Yaghan, 2008). These examples show that current safety evaluation frameworks that predominantly evaluate languages in a monolingual setting fail to capture the complex reality of multilingual communication. Future work on multilingual red-teaming should develop a methodology that systematically accounts for diverse multilingual multi-turn interactions among users (Li et al., 2025) to ensure that models remain safe across the full spectrum of real-world usage patterns rather than just in artificial monolingual test scenarios.

3.2 Culturally-Contextualized Synthetic Training Data

Collecting labeled training data for LLM safety alignment can be resource-intensive, and many English-centric research has turned to using synthetic data generation (Bai et al., 2022; Kruschwitz and Schmidhuber, 2024; Samvelyan et al., 2024). However, exploration of multilingual synthetic safety data has been relatively underexplored. Here, we propose two viable future research directions based on *constitutional AI* framework (Bai et al., 2022; Kundu et al., 2023).

LLM Generation. Under constitutional AI framework, LLMs are first prompted to generate harmful (or harmless texts). They are then presented with a set of human-written principles that capture culture-specific harms so that they can engage in a multi-turn process of critiquing and revising originally harmful generations to harmless generations (or vice versa), to create culture-specific preference pairs for alignment training. Enabling constitutional AI for multilingual and multicultural alignment data generation requires close collaboration among linguists, cultural anthropologists and AI researchers to co-create three key components: (1) culturally-informed constitutional principles that reflect diverse value systems and ethical frameworks across different societies (Kirk et al., 2024; Pistilli et al., 2025); (2) sufficiently capable multilingual LLMs that can both understand these principles and generate high-quality content in target languages (Qin et al., 2024; Huang et al., 2024a); and (3) evaluation protocols involving native speakers and cultural experts to validate both the constitutional principles and the resulting synthetic data (Kyrychenko et al., 2025). This direction offers a pathway toward scalable, culturally grounded alignment practices that make LLM safety more inclusive and globally relevant.

Machine Translation. Machine translation (MT) often fails to capture or preserve culture-specific harms and may introduce undesirable societal biases such as gender stereotyping (Savoldi et al., 2021; Ahn et al., 2022; Wang et al., 2022; Costa-jussà et al., 2023b,c). The iterative refinement process from the constitutional AI framework can detect and mitigate translation artifacts that might inadvertently encode harmful content or lose important cultural nuances. Unlike direct LLM generation, this approach can take advantage of the decades-long research in MT, especially on cross-cultural adaptation studies (Maxwell et al., 1996; de Lima Barroso et al., 2018; Gorecki et al., 2014; Mbada et al., 2015; Pilz et al., 2014). Future work should focus on developing automated methods to identify culture-specific safety issues that might be lost in translation, especially for languages with limited digital presence and linguistic resources.

3.3 Towards Understanding Crosslingual Safety Generalization

Most existing safety alignment data are centered on English or Chinese (Röttger et al., 2025; Costa-jussà et al., 2024; Plaza-del Arco et al., 2024). It is important to understand how safety alignment generalizes across languages, so the model developers can *anticipate* potential failure modes when alignment training data lack language coverage.

Mechanistic interpretability. This scientific approach of reverse-engineering neural networks to understand precisely how they process information at the circuit and component levels——allows researchers to characterize mechanisms that enable or prevent safety alignment knowledge transfer. We believe this research direction is particularly helpful in explaining several phenomena, such as why detoxification and debiasing can transfer effectively across languages (Li et al., 2024; Reusens et al., 2023) but not refusal training (Shen et al., 2024; Aakanksha et al., 2024; Wang et al., 2025), or to what extent safety alignment is *preserved* after language adaptation to underrepresented languages (Yong et al., 2023b; Lin et al., 2024; Ji et al., 2024). Insights from this research direction can inspire novel training techniques that facilitate zero-shot crosslingual generalization of alignment training and maintain safety consistency as language coverage expands.

Training data influence analysis. We also recommend exploring the use of influence functions (Grosse et al., 2023; Ruis et al., 2025) to study crosslingual alignment. This technique enables researchers to trace how specific training examples causally affect model behavior during generation. Training data influence analysis offers a valuable complement to mechanistic approaches for investigating two key open questions. For crosslingual generalization, it can help quantify how safety-relevant examples—especially those from high-resource versus low-resource languages—contribute to harmful or aligned outputs. For language adaptation, influence functions can identify problematic documents within the continued pretraining corpus, enabling more targeted curation of safer language-specific data. To our knowledge, there is currently *very limited work* on analyzing training-example-to-output relationships for multilingual safety-relevant behaviors. This presents a promising and underexplored direction for improving alignment practices across languages.

4 Related Work and Discussion

Our work contrasts prior survey literature on multilingual NLP (Joshi et al., 2020; Pamungkas et al., 2023; Yadav and Sitaram, 2022; Winata et al., 2023; Huang et al., 2024a; Qin et al., 2024; Wu et al., 2025) by focusing on LLM safety. The limitations we identify align with concerns by Blasi et al. (2022) regarding systematic inequalities in language technology, which privileges certain sociolinguistic groups through choices in data collection, annotation protocols, and evaluation. Our findings suggest these inequalities may be even more pronounced in safety research, where cultural and linguistic nuances significantly impact harm and mitigation strategies.

Recent efforts to catalog LLM safety research challenges (Barez et al., 2025; Debar et al., 2024; Anwar et al., 2024) have primarily centered on threats identified through English-language models, often overlooking multilingual aspects. This gap, along with our survey findings, echoes the "square-one bias" phenomenon (Ruder et al., 2022): When NLP researchers moves beyond optimizing for usefulness (e.g., accuracy), their study is often only conducted in a single direction of either safety, interpretability, or multilinguality. This siloed approach means that progress in one dimension rarely informs the others, resulting in a fragmented research landscape where multilingual LLM safety research remains underdeveloped.

5 Conclusion

Our analysis of nearly 300 publications (2020-2024) reveals a significant language gap in LLM safety research, with even high-resource non-English languages receiving minimal attention and typically appearing only in multi-language studies that lack the depth of English-focused work. This linguistic imbalance potentially leaves language-specific risks undetected as LLMs deploy globally. To address these disparities, we make recommendations to future conferences and highlight several critical future research directions.

Limitations

Coverage of venues Due to the focus on *ACL venues, we might have missed out on relevant multilingual safety works that are either not peer reviewed (yet) or published in other venues, such as ML conferences and workshops. Since it is a very fast moving field, the state of the field described in this paper represents a snapshot in time. We hope that if we ran an analysis like this in a year's time, the data would hopefully paint a more optimistic picture.

Annotation accuracy Inaccuracies in our annotations might have introduced imprecision in our measurements of the language gap. From our analysis of the inter-annotator agreement, we suspect that this would foremost affect the categorization of safety research topics, as the labels for these categories carry the most ambiguity. When papers do not state language coverage very prominently, such in the abstract or introduction or the experimental setup, it might lead to oversight in the annotation (reducing recall in annotations), depending how deeply an annotator reads the paper. However, we observe that especially those works that are investigating multilinguality in LLM safety as a primary angle, do state it explicitly, so we are confident we did not miss these.

Research directions We highlight three prominent future directions for multilingual safety research in our work, but we believe there are many other directions that are equally important for advancing safety and security of LLMs in global deployment. These include work on AI governance, hate speech detection, multimodal AI safety, algorithmic designs for multilingual alignment training, etc. Fundamentally, our work illuminates the substantial language disparity within current LLM safety research. Therefore, as researchers pursue diverse research directions on LLM safety, efforts on bridging this linguistic divide must remain central to ensuring equitable safeguards across the world's languages.

Acknowledgements and Disclosure

We are grateful for feedback from Hellina Hailu Nigatu, M Saiful Bari, Cristina Menghini, Alham Fikri Aji, Pedro Ortiz Suarez, Victor Ojewale, Simran Khanuja, Arianna Muti, Debora Nozza, Srishti Yadav, Jonas Kgomo, and Catherine Arnett on the early draft of our work. Disclosure: Stephen Bach is an advisor to Snorkel AI, a company that provides software and services for datacentric artificial intelligence.

References

Aakanksha, Arash Ahmadian, Beyza Ermis, Seraphina Goldfarb-Tarrant, Julia Kreutzer, Marzieh Fadaee, and Sara Hooker. 2024. The multilingual alignment prism: Aligning global and local preferences to reduce harm. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12027–12049, Miami, Florida, USA. Association for Computational Linguistics.

Jaimeen Ahn, Hwaran Lee, Jinhwa Kim, and Alice Oh. 2022. Why knowledge distillation amplifies gender bias and how to mitigate from the perspective of DistilBERT. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 266–272, Seattle, Washington. Association for Computational Linguistics.

Mansour Al Ghanim, Saleh Almohaimeed, Mengxin Zheng, Yan Solihin, and Qian Lou. 2024. Jailbreaking LLMs with Arabic transliteration and Arabizi. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18584–18600, Miami, Florida, USA. Association for Computational Linguistics.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, and 1 others. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, and 1 others. 2024. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Patricia Baquedano-López and Shlomy Kattan. 2007. Growing up in a multilingual community: Insights from language socialization. *Handbook of multilingualism and multilingual communication*, 5:69–99.

Fazl Barez, Tingchen Fu, Ameya Prabhu, Stephen Casper, Amartya Sanyal, Adel Bibi, Aidan O'Gara, Robert Kirk, Ben Bucknall, Tim Fist, and 1 others. 2025. Open problems in machine unlearning for ai safety. *arXiv preprint arXiv:2501.04952*.

Dipto Barman, Ziyi Guo, and Owen Conlan. 2024. The dark side of language models: Exploring the potential of llms in multimedia disinformation generation and dissemination. *Machine Learning with Applications*, page 100545.

Emily Bender. 2019. The #benderrule: On naming the languages we study and why it matters. *The Gradient*.

Emily M Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6.

Yoshua Bengio, Sören Mindermann, and Daniel Privitera. 2025. International ai safety report 2025.

Mukul Bhutani, Kevin Robinson, Vinodkumar Prabhakaran, Shachi Dave, and Sunipa Dev. 2024. SeeGULL multilingual: a dataset of geo-culturally situated stereotypes. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 842–854, Bangkok, Thailand. Association for Computational Linguistics.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explains the cross-lingual capabilities of English pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3563–3574, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lorna Carson and Ning Jiang. 2021. Offensive words in chinese dialects. In *An Anatomy of Chinese Offensive Words: A Lexical and Semantic Analysis*, pages 99–143. Springer.

Yik Siu Chan, Narutatsu Ri, Yuxin Xiao, and Marzyeh Ghassemi. 2025. Speak easy: Eliciting harmful jailbreaks from llms with simple interactions. *arXiv preprint arXiv:2502.04322*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Team Cohere, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, and 1 others. 2025. Command a: An enterprise-ready large language model. *arXiv preprint arXiv:2504.00698*.

Marta Costa-jussà, Pierre Andrews, Christine Basta, Juan Ciro, Agnieszka Falenska, Seraphina Goldfarb-Tarrant, Rafael Mosquera, Debora Nozza, and Eduardo Sánchez. 2024. Overview of the shared task on machine translation gender bias evaluation with multilingual holistic bias. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 399–404, Bangkok, Thailand. Association for Computational Linguistics.

Marta Costa-jussà, Pierre Andrews, Eric Smith, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Daniel Licht, and Carleigh Wood. 2023a. Multilingual holistic bias: Extending descriptors and patterns to unveil demographic biases in languages at scale. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14141–14156, Singapore. Association for Computational Linguistics.

Marta Costa-jussà, Pierre Andrews, Eric Smith, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Daniel Licht, and Carleigh Wood. 2023b. Multilingual holistic bias: Extending descriptors and patterns to unveil demographic biases in languages at scale. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14141–14156, Singapore. Association for Computational Linguistics.

Marta Costa-jussà, Eric Smith, Christophe Ropers, Daniel Licht, Jean Maillard, Javier Ferrando, and Carlos Escolano. 2023c. Toxicity in multilingual machine translation at scale. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9570–9586, Singapore. Association for Computational Linguistics.

Tianyu Cui, Yanling Wang, Chuanpu Fu, Yong Xiao, Sijia Li, Xinhao Deng, Yunpeng Liu, Qinglin Zhang, Ziyi Qiu, Peiyang Li, and 1 others. 2024. Risk taxonomy, mitigation, and assessment benchmarks of large language model systems. *arXiv preprint arXiv:2401.05778*.

Bárbara Iansã de Lima Barroso, Cláudia Regina Cabral Galvão, Luiz Bueno da Silva, and Selma Lancman.

2018. A systematic review of translation and cross-cultural adaptation of instruments for the selection of assistive technologies. *Occupational Therapy International*, 2018(1):4984170.

Herve Debar, Sven Dietrich, Pavel Laskov, Emil C Lupu, and Eirini Ntoutsi. 2024. Emerging security challenges of large language models. *arXiv preprint arXiv:2412.17614*.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2024. Multilingual jailbreak challenges in large language models. In *The Twelfth International Conference on Learning Representations*.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2024. We don't talk about that: Case studies on intersectional analysis of social bias in large language models. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 33–44, Bangkok, Thailand. Association for Computational Linguistics.

Jean-Marc Dewaele. 2004. The emotional force of swearwords and taboo words in the speech of multilinguals. *Journal of multilingual and multicultural development*, 25(2-3):204–222.

Zhichen Dong, Zhanhui Zhou, Chao Yang, Jing Shao, and Yu Qiao. 2024. Attacks, defenses and evaluations for LLM conversation safety: A survey. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6734–6747, Mexico City, Mexico. Association for Computational Linguistics.

Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, and Wei Xu. 2024. Reducing privacy risks in online self-disclosures with language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13732–13754, Bangkok, Thailand. Association for Computational Linguistics.

Beyza Ermis, Luiza Pozzobon, Sara Hooker, and Patrick Lewis. 2024. From one to many: Expanding the scope of toxicity mitigation in language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15041–15058, Bangkok, Thailand. Association for Computational Linguistics.

Daniel L Everett. 2012. *Language: The cultural tool*. Vintage.

Mahmud Fasya and Dini Gilang Sari. 2021. Socio-cultural factors that determine language choice in a multilingual society. In *Fifth International Conference on Language, Literature, Culture, and Education (ICOLLITE 2021)*, pages 412–418. Atlantis Press.

Saadia Gabriel, Isha Puri, Xuhai Xu, Matteo Malgaroli, and Marzyeh Ghassemi. 2024. Can AI relate: Testing large language model response for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2206–2221, Miami, Florida, USA. Association for Computational Linguistics.

Penelope Gardner-Chloros. 2009. *Code-switching*. Cambridge university press.

Catalina Goanta, Nikolaos Aletras, Ilias Chalkidis, Sofia Ranchordás, and Gerasimos Spanakis. 2023. Regulation and NLP (RegNLP): Taming large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8712–8724, Singapore. Association for Computational Linguistics.

Claudia Gorecki, Julia M Brown, Michelle Briggs, Suzanne Coleman, Carol Dealey, Elizabeth McGinnis, E Andrea Nelson, Nikki Stubbs, Lyn Wilson, and Jane Nixon. 2014. Language translation & cross-cultural adaptation guideline. *Recommendations for language translation and cross-cultural adaption of the PU-QOL questionnaire*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, and 1 others. 2023. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*.

Melody Y Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, and 1 others. 2024. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*.

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Hasan Abed Al Kader Hammoud, Umberto Michieli, Fabio Pizzati, Philip Torr, Adel Bibi, Bernard Ghanem, and Mete Ozay. 2024. Model merging and safety alignment: One bad model spoils the bunch. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13033–13046, Miami, Florida, USA. Association for Computational Linguistics.

Sabit Hassan, Anthony Sicilia, and Malihe Alikhani. 2024. Active learning for robust and representative LLM generation in safety-critical scenarios. In *Proceedings of the 1st Workshop on Customizable NLP:*

Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U), pages 113–123, Miami, Florida, USA. Association for Computational Linguistics.

Xuanli He, Jun Wang, Qiongkai Xu, Pasquale Minervini, Pontus Stenetorp, Benjamin IP Rubinstein, and Trevor Cohn. 2024. Tuba: Cross-lingual transferability of backdoor attacks in llms with instruction tuning. *arXiv preprint arXiv:2404.19597*.

Harry Ed Hoijer. 1954. Language in culture; conference on the interrelations of language and other aspects of culture.

Wenyue Hua, Xianjun Yang, Mingyu Jin, Zelong Li, Wei Cheng, Ruixiang Tang, and Yongfeng Zhang. 2024. TrustAgent: Towards safe and trustworthy LLM-based agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10000–10016, Miami, Florida, USA. Association for Computational Linguistics.

Kaiyu Huang, Fengran Mo, Xinyu Zhang, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, and 1 others. 2024a. A survey on large language models with multilingualism: Recent advances and new frontiers. *arXiv preprint arXiv:2405.10936*.

Xin Huang, Tarun Kumar Vangani, Minh Duc Pham, Xunlong Zou, Bin Wang, Zhengyuan Liu, and Ai Ti Aw. 2024b. Meralion-textllm: Cross-lingual understanding of large language models in chinese, indonesian, malay, and singlish. *arXiv preprint arXiv:2501.08335*.

Devansh Jain, Priyanshu Kumar, Samuel Gehman, Xuhui Zhou, Thomas Hartvigsen, and Maarten Sap. 2024. Polyglotoxicityprompts: Multilingual evaluation of neural toxic degeneration in large language models. In *First Conference on Language Modeling*.

Shaoxiong Ji, Zihao Li, Indraneil Paul, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayyán O'Brien, Hengyu Luo, Hinrich Schütze, Jörg Tiedemann, and 1 others. 2024. Emma-500: Enhancing massively multilingual adaptation of large language models. *arXiv preprint arXiv:2409.17892*.

Wenying Jiang. 2000. The relationship between culture and language. *ELT journal*, 54(4):328–334.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Arturs Kanepajs, Vladimir Ivanov, and Richard Moulange. 2024. Towards safe multilingual frontier AI. In *Workshop on Socially Responsible Language Modelling Research*.

Teo Keipi, Matti Näsi, Atte Oksanen, and Pekka Räsänen. 2016. *Online hate and harmful content: Cross-national perspectives*. Taylor & Francis.

Tseen-Ling Khoo. 2003. *Banana Bending: Asian-Australian and Asian-Canadian Literatures*. Hong Kong University Press.

Paria Khoshtab, Danial Namazifard, Mostafa Masoudi, Ali Akhgary, Samin Mahdizadeh Sani, and Yadollah Yaghoobzadeh. 2025. Comparative study of multilingual idioms and similes in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8680–8698, Abu Dhabi, UAE. Association for Computational Linguistics.

Minbeom Kim, Jahyun Koo, Hwanhee Lee, Joonsuk Park, Hwaran Lee, and Kyomin Jung. 2024. LifeTox: Unveiling implicit toxicity in life advice. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 688–698, Mexico City, Mexico. Association for Computational Linguistics.

Hannah Rose Kirk, Alexander Whitefield, Paul Rottger, Andrew M Bean, Katerina Margatina, Rafael Mosquera-Gomez, Juan Ciro, Max Bartolo, Adina Williams, He He, and 1 others. 2024. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37:105236–105344.

Katerina Korre, Arianna Muti, Federico Ruggeri, and Alberto Barrón-Cedeño. 2025. Untangling hate speech definitions: A semantic componential analysis across cultures and domains. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3184–3198, Albuquerque, New Mexico. Association for Computational Linguistics.

Claire Kramsch. 2014. Language and culture. *AILA review*, 27(1):30–55.

Julia Kreutzer, Eleftheria Briakou, Sweta Agrawal, Marzieh Fadaee, and Kocmi Tom. 2025. D\'ej\a vu: Multilingual llm evaluation through the lens of machine translation evaluation. *arXiv preprint arXiv:2504.11829*.

Udo Kruschwitz and Maximilian Schmidhuber. 2024. LLM-based synthetic datasets: Applications and limitations in toxicity detection. In *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024*, pages 37–51, Torino, Italia. ELRA and ICCL.

Sandipan Kundu, Yuntao Bai, Saurav Kadavath, Amanda Askell, Andrew Callahan, Anna Chen, Anna Goldie, Avital Balwit, Azalia Mirhoseini, Brayden McLean, and 1 others. 2023. Specific versus general principles for constitutional ai. *arXiv preprint arXiv:2310.13798*.

Yara Kyrychenko, Ke Zhou, Edyta Bogucka, and Daniele Quercia. 2025. C3ai: Crafting and evaluating constitutions for constitutional ai. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 3204–3218, New York, NY, USA. Association for Computing Machinery.

Benjamin Larsen and Virginia Dignum. 2024. Ai value alignment: How we can align artificial intelligence with human values. *World Economic Forum*.

Xiaochen Li, Zheng Xin Yong, and Stephen Bach. 2024. Preference tuning for toxicity mitigation generalizes across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13422–13440, Miami, Florida, USA. Association for Computational Linguistics.

Yubo Li, Xiaobin Shen, Xinyu Yao, Xueying Ding, Yidi Miao, Ramayya Krishnan, and Rema Padman. 2025. Beyond single-turn: A survey on multi-turn interactions with large language models. *arXiv preprint arXiv:2504.04717*.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André FT Martins, and Hinrich Schütze. 2024. Mala-500: Massive language adaptation of large language models. *arXiv preprint arXiv:2401.13303*.

Edoardo Manino, Julia Rozanova, Danilo Carvalho, Andre Freitas, and Lucas Cordeiro. 2022. Systematicity, compositionality and transitivity of deep NLP models: a metamorphic testing perspective. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2355–2366, Dublin, Ireland. Association for Computational Linguistics.

Beverley Maxwell, MO Martin, and DL Kelly. 1996. Translation and cultural adaptation of the survey instruments. *Third international mathematics and science study (TIMSS) technical report*, 1:159–169.

Abdelfattah Mazari and Naoual Derraz. 2015. Language and culture. *International Journal of Humanities and Cultural Studies*, 2(2):350–359.

Chidozie Emmanuel Mbada, Gafar Atanda Adeogun, Michael Opeoluwa Ogunlana, Rufus Adesoji Adedoyin, Adesanmi Akinsulore, Taofeek Oluwole Awotidebe, Opeyemi Ayodiipo Idowu, and Olumide Ayoola Olaoye. 2015. Translation, cross-cultural adaptation and psychometric evaluation of yoruba version of the short-form 36 health survey. *Health and quality of life outcomes*, 13:1–12.

Preslav Nakov, Firoj Alam, Shaden Shaar, Giovanni Da San Martino, and Yifan Zhang. 2021. COVID-19 in Bulgarian social media: Factuality, harmfulness, propaganda, and framing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 997–1009, Held Online. INCOMA Ltd.

Zabir Al Nazi and Wei Peng. 2024. Large language models in healthcare and medical domain: A review. In *Informatics*, volume 11, page 57. MDPI.

Alexander Tobias Neumann, Yue Yin, Sulayman Sowe, Stefan Decker, and Matthias Jarke. 2024. An llm-driven chatbot in higher education for databases and information systems. *IEEE Transactions on Education*.

Hellina Hailu Nigatu and Inioluwa Deborah Raji. 2024. "i searched for a religious song in amharic and got sexual content instead": Investigating online harm in low-resourced languages on youtube. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 141–160.

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

Chad Nilep. 2006. "code switching" in sociocultural linguistics. *Colorado research in linguistics*.

Nobal B. Niraula, Saurab Dulal, and Diwa Koirala. 2021. Offensive language detection in Nepali social media. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, Online. Association for Computational Linguistics.

OpenAI. 2022. Introducing ChatGPT. `https://openai.com/blog/chatgpt/`. Accessed [Insert Date You Accessed This Page].

OpenAI. 2025. Openai gpt-4.5 system card. Technical report, OpenAI.

Ankit Pal and Malaikannan Sankarasubbu. 2024. Gemini goes to Med school: Exploring the capabilities of multimodal large language models on medical challenge problems & hallucinations. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 21–46, Mexico City, Mexico. Association for Computational Linguistics.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2023. Towards multidomain and multilingual abusive language detection: a survey. *Personal and Ubiquitous Computing*, 27(1):17–43.

Bruna Pilz, Rodrigo A Vasconcelos, Freddy B Marcondes, Samuel S Lodovichi, Wilson Mello, and Débora B Grossi. 2014. The brazilian version of start back screening tool-translation, cross-cultural adaptation and reliability. *Brazilian journal of physical therapy*, 18:453–461.

Giada Pistilli, Alina Leidinger, Yacine Jernite, Atoosa Kasirzadeh, Alexandra Sasha Luccioni, and Margaret Mitchell. 2025. Civics: Building a dataset for examining culturally-informed values in large language models. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '24, page 1132–1144. AAAI Press.

Flor Miriam Plaza-del Arco, Debora Nozza, Marco Guerini, Jeffrey Sorensen, and Marcos Zampieri. 2024. Countering hateful and offensive speech online - open challenges. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 11–16, Miami, Florida, USA. Association for Computational Linguistics.

Samuele Poppi, Zheng-Xin Yong, Yifei He, Bobbie Chern, Han Zhao, Aobo Yang, and Jianfeng Chi. 2025. Towards understanding the fragility of multilingual llms against fine-tuning attacks. *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. 2024. Multilingual large language model: A survey of resources, taxonomy and frontiers. *arXiv preprint arXiv:2404.04925*.

Manon Reusens, Philipp Borchert, Margot Mieskes, Jochen De Weerdt, and Bart Baesens. 2023. Investigating bias in multilingual language models: Cross-lingual transfer of debiasing techniques. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2887–2896, Singapore. Association for Computational Linguistics.

Paul Röttger, Fabio Pernisi, Bertie Vidgen, and Dirk Hovy. 2025. Safetyprompts: a systematic review of open datasets for evaluating and improving large language model safety. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27617–27627.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2022. Square one bias in NLP: Towards a multi-dimensional exploration of the research manifold. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2340–2354, Dublin, Ireland. Association for Computational Linguistics.

Laura Ruis, Maximilian Mozes, Juhan Bae, Siddhartha Rao Kamalakara, Dwaraknath Gnaneshwar, Acyr Locatelli, Robert Kirk, Tim Rocktäschel, Edward Grefenstette, and Max Bartolo. 2025. Procedural knowledge in pretraining drives reasoning in large language models. In *The Thirteenth International Conference on Learning Representations*.

Mikayel Samvelyan, Sharath Chandra Raparthy, Andrei Lupu, Eric Hambro, Aram Markosyan, Manish Bhatt, Yuning Mao, Minqi Jiang, Jack Parker-Holder, Jakob Foerster, and 1 others. 2024. Rainbow teaming: Open-ended generation of diverse adversarial prompts. *Advances in Neural Information Processing Systems*, 37:69747–69786.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Ayse Pinar Saygin. 2001. Processing figurative language in a multi-lingual task: Translation, transfer and metaphor. In *Proceedings of the Workshop on Corpus-based and Processing Approaches to Figurative Language*. Citeseer.

Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. 2024. The language barrier: Dissecting safety challenges of LLMs in multilingual contexts. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2668–2680, Bangkok, Thailand. Association for Computational Linguistics.

Dan Shi, Tianhao Shen, Yufei Huang, Zhigen Li, Yongqi Leng, Renren Jin, Chuang Liu, Xinwei Wu, Zishan Guo, Linhao Yu, Ling Shi, Bojian Jiang, and Deyi Xiong. 2024a. Large language model safety: A holistic survey. *Preprint*, arXiv:2412.17686.

Shaojie Shi, Xiaoyu Tan, Xihe Qiu, Chao Qu, Kexin Nie, Yuan Cheng, Wei Chu, Xu Yinghui, and Yuan Qi. 2024b. ULMR: Unlearning large language models via negative response and model parameter average. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 755–762, Miami, Florida, US. Association for Computational Linguistics.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8.

Jiayang Song, Yuheng Huang, Zhehua Zhou, and Lei Ma. 2025. Multilingual blending: Large language model safety alignment evaluation with language mixture. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3433–3449, Albuquerque, New Mexico. Association for Computational Linguistics.

Kamal K Sridhar. 1996. Societal multilingualism. *Sociolinguistics and language teaching*, 47:70.

Alex Tamkin, Miles McCain, Kunal Handa, Esin Durmus, Liane Lovitt, Ankur Rathi, Saffron Huang, Alfred Mountfield, Jerry Hong, Stuart Ritchie, and 1 others. 2024. Clio: Privacy-preserving insights into real-world ai use. *arXiv preprint arXiv:2412.13678*.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay

Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Monica M Trieu. 2019. Understanding the use of "twinkie,""banana," and "fob": Identifying the origin, role, and consequences of internalized racism within asian america. *Sociology Compass*, 13(5):e12679.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Jun Wang, Benjamin Rubinstein, and Trevor Cohn. 2022. Measuring and mitigating name biases in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2576–2590, Dublin, Ireland. Association for Computational Linguistics.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael Lyu. 2024a. All languages matter: On the multilingual safety of LLMs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5865–5877, Bangkok, Thailand. Association for Computational Linguistics.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael Lyu. 2024b. All languages matter: On the multilingual safety of LLMs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5865–5877, Bangkok, Thailand. Association for Computational Linguistics.

Xiangwen Wang, Jie Peng, Kaidi Xu, Huaxiu Yao, and Tianlong Chen. 2024c. Reinforcement learning-driven LLM agent for automated attacks on LLMs. In *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, pages 170–177, Bangkok, Thailand. Association for Computational Linguistics.

Xinpeng Wang, Mingyang Wang, Yihong Liu, Hinrich Schütze, and Barbara Plank. 2025. Refusal direction is universal across safety-aligned languages. *arXiv preprint arXiv:2505.17306*.

Yixu Wang, Yan Teng, Kexin Huang, Chengqi Lyu, Songyang Zhang, Wenwei Zhang, Xingjun Ma, Yu-Gang Jiang, Yu Qiao, and Yingchun Wang. 2024d. Fake alignment: Are LLMs really aligned well? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4696–4712, Mexico City, Mexico. Association for Computational Linguistics.

Zhonghao Wang, Zijia Lu, Bo Jin, and Haiying Deng. 2023. Mediagpt: A large language model for chinese media. *arXiv preprint arXiv:2307.10930*.

Qingsong Wen, Jing Liang, Carles Sierra, Rose Luckin, Richard Tong, Zitao Liu, Peng Cui, and Jiliang Tang. 2024. Ai for education (ai4edu): Advancing personalized education with llm and adaptive learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6743–6744.

Genta Winata, Alham Fikri Aji, Zheng Xin Yong, and Thamar Solorio. 2023. The decades progress on code-switching research in NLP: A systematic survey on trends and challenges. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978, Toronto, Canada. Association for Computational Linguistics.

Minghao Wu, Weixuan Wang, Sinuo Liu, Huifeng Yin, Xintong Wang, Yu Zhao, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. The bitter lesson learned from 2,000+ multilingual benchmarks. *arXiv preprint arXiv:2504.15521*.

Hemant Yadav and Sunayana Sitaram. 2022. A survey of multilingual models for automatic speech recognition. *arXiv preprint arXiv:2202.12576*.

Mohammad Ali Yaghan. 2008. " arabizi": A contemporary style of arabic slang. *Design issues*, 24(2):39–52.

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

Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. 2024b. Self-distillation bridges distribution gap in language model fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1028–1043, Bangkok, Thailand. Association for Computational Linguistics.

Zheng Xin Yong, Cristina Menghini, and Stephen Bach. 2023a. Low-resource languages jailbreak GPT-4. In *Socially Responsible Language Modelling Research*.

Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. 2023b. BLOOM+1: Adding language support to BLOOM for zero-shot prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11682–11703, Toronto, Canada. Association for Computational Linguistics.

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

Tongxin Yuan, Zhiwei He, Lingzhong Dong, Yiming Wang, Ruijie Zhao, Tian Xia, Lizhen Xu, Binglin Zhou, Fangqi Li, Zhuosheng Zhang, Rui Wang, and Gongshen Liu. 2024. R-judge: Benchmarking safety risk awareness for LLM agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1467–1490, Miami, Florida, USA. Association for Computational Linguistics.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024. How johnny can persuade LLMs to jailbreak them: Rethinking persuasion to challenge AI safety by humanizing LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14322–14350, Bangkok, Thailand. Association for Computational Linguistics.

Yizhou Zhang, Karishma Sharma, Lun Du, and Yan Liu. 2024a. Toward mitigating misinformation and social media manipulation in llm era. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 1302–1305, New York, NY, USA. Association for Computing Machinery.

Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, and 1 others. 2024b. Simulating classroom education with llm-empowered agents. *arXiv preprint arXiv:2406.19226*.

Zhanhui Zhou, Jie Liu, Jing Shao, Xiangyu Yue, Chao Yang, Wanli Ouyang, and Yu Qiao. 2024. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10586–10613, Bangkok, Thailand. Association for Computational Linguistics.

Shucheng Zhu, Bingjie Du, Jishun Zhao, Ying Liu, and Pengyuan Liu. 2024. Do PLMs and annotators share the same gender bias? definition, dataset, and framework of contextualized gender bias. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 20–32, Bangkok, Thailand. Association for Computational Linguistics.