Evaluating Multilingual Safety Benchmarks for Low-Resource Languages in the Majority World

Alisar Mustafa and Cherry Wu

As large language models are deployed across multilingual environments, the benchmarks used to evaluate their safety remain largely designed for high-resource, English-dominant contexts. Trust & Safety systems increasingly rely on automated classifiers and generative models to moderate harmful content across dozens of languages, yet the tools used to assess them often fail to capture linguistic variation and culturally specific harm. This paper presents a structured review of 37 benchmarks relevant to multilingual safety evaluation, 27 of them safety benchmarks, published between 2019 and 2026. Using a nine-dimension taxonomy spanning linguistic authenticity, cultural grounding, and evaluation transparency, we analyze how benchmarks construct evaluation datasets and report safety performance. We identify recurring design patterns: reliance on translated English prompts rather than native-language data, aggregate metrics that obscure cross-language variation, and limited use of locally grounded harm categories or community-informed evaluation. These patterns show that broader language coverage alone does not ensure benchmarks capture how harm is expressed across cultural contexts. We therefore propose a design framework emphasizing native-authored data, disaggregated reporting, locally grounded harm taxonomies, symmetric evaluation of false positives, coverage of dialect and code-switching, and testing under deployment-relevant conditions.

1 Introduction

Large language models (LLMs) are now deployed at scale across Majority World contexts, including South Asia, sub-Saharan Africa, Southeast Asia, and Latin America, where multilingual communication and dialect mixing are common features of everyday language use. The benchmarks used to evaluate the safety of these systems were largely developed under different linguistic and policy assumptions. Most benchmarks originated in

Journal of Online Trust and Safety 3 (3)
DOI: 10.54501/jots.v3i3.330
Research Article

Stanford Tech Impact and Policy Center

English-language settings, were later extended to other languages primarily through translation, and were validated against harm taxonomies shaped by Western regulatory and policy environments. Prior research has documented a persistent "language gap" in AI safety research, where English-language datasets and evaluations dominate while multilingual contexts remain underexplored (Yong et al. 2025; Peppin et al. 2025). As a result, the ability of current benchmarks to capture risks arising in diverse sociolinguistic contexts remains uneven.

Here we survey 37 benchmarks relevant to multilingual safety evaluation, 27 of them safety benchmarks, published between 2019 and 2026. These benchmarks evaluate models whose safety training is largely English-centric. Techniques such as reinforcement learning from human feedback, Constitutional AI, and instruction tuning are implemented predominantly using English-language datasets, so the multilingual capabilities acquired during pretraining coexist with more limited multilingual safety calibration (Yue Deng et al. 2024; Kumar et al. 2025).

This review provides two primary contributions. First, it provides a taxonomy-based structured review of 37 benchmarks published between 2019 and 2026, comprising 27 multilingual and monolingual safety benchmarks alongside 10 capability benchmarks analyzed as methodological reference points. Using a nine-dimension framework spanning linguistic authenticity, cultural grounding, evaluation transparency, and operational readiness, the review produces a comparative map of gaps in current benchmark design. The taxonomy enables systematic comparison across benchmarks with heterogeneous methods and objectives. Second, it proposes a six-principle design framework for multilingual safety evaluation. The framework draws directly from patterns observed in the benchmark corpus (each principle corresponds to a findings subsection in Section 4) and is intended to provide practical guidance for researchers and practitioners developing evaluation infrastructure.

2 Background

2.1 Safety Alignment and English-Dominant Training Pipelines

The safety properties of current LLMs reflect deliberate alignment procedures, including reinforcement learning from human feedback (RLHF), Constitutional AI, and instruction tuning. In practice, however, much of the data and evaluation infrastructure used for safety alignment remains concentrated in English-language contexts. As a result, multilingual capabilities acquired during pretraining on diverse corpora coexist with safety calibration that is largely English-centric (Yong et al. 2025; Peppin et al. 2025).

Evidence consistent with this asymmetry appears across the literature. Multilingual capabilities typically emerge during the pretraining stage, while safety fine-tuning relies on more limited multilingual coverage in later stages (Yue Deng et al. 2024). Kumar et

al. (2025) note a related constraint in their own system, observing higher rates of unsafe responses in several non-English languages. As a result, tools developed to address multilingual safety challenges often inherit the training distributions of the datasets on which they are built.

Empirical patterns associated with this imbalance appear across multiple studies. S-Eval reports declining safety performance along a language-resource gradient, from 55.85% in Chinese to 46.15% in Swahili (Yuan et al. 2025). MultiJail similarly reports that low-resource languages produce harmful outputs at approximately three times the rate observed in high-resource languages (Yue Deng et al. 2024). The recurrence of these findings across independent studies suggests that disparities in safety performance may be linked to characteristics of current training and evaluation pipelines.

These observations have implications for benchmark design. Benchmarks limited to high-resource languages, or extended to additional languages primarily through translation from English, largely measure the extent to which safety behaviors learned in high-resource contexts transfer across languages. Such evaluations provide only partial visibility into safety performance in the broader linguistic environments where models are deployed. Assessing safety in low-resource contexts requires evaluation frameworks designed for those contexts from the outset.

2.2 Conceptualizing “Low-Resource” Languages

The term low-resource is used inconsistently across the NLP literature and often conflates distinct conditions. For this paper, we distinguish three related but analytically separate forms of scarcity.

Data scarcity refers to limited pretraining and fine-tuning data in a given language, typically measured through indicators such as CommonCrawl corpus proportions or Wikipedia article counts. Research attention and data resources remain concentrated on a small number of languages (P. Joshi et al. 2020). This definition underlies benchmark categorizations in work such as MultiJail (Yue Deng et al. 2024) and LinguaSafe (Ning et al. 2025). Bengali, for example, is commonly classified as low-resource despite having approximately 230 million native speakers, reflecting gaps in digital data infrastructure rather than speaker population.

Evaluation scarcity refers to the absence of benchmarks designed for a particular language. A language may have modest representation in training data yet still possess well-developed evaluation infrastructure, or conversely be widely spoken while lacking purpose-built safety benchmarks. This paper focuses primarily on evaluation scarcity. Within the 37 works in our corpus, languages with the most extensive native-authored evaluation coverage tend to coincide with dedicated institutional investment, including Arabic (AraSafe; Mubarak, Mohamed, and Hawasly 2025), Indonesian (IndoSafety; Azmi et al. 2025), and Bengali (BanglaGuard; Alam et al. 2026; BengaliMoralBench; Ridoy

et al. 2026). Large linguistic regions, including Central Asia, Indigenous languages of the Americas, and much of sub-Saharan Africa, remain largely absent from current safety evaluation efforts.

Institutional scarcity refers to limited research infrastructure, annotation capacity, and community representation in benchmark development. Building benchmarks for low-resource languages frequently requires multi-institutional collaboration and external funding, as documented across several works in the corpus (Muhammad et al. 2025; Ridoy et al. 2026; Wynter et al. 2025). Section 4 examines these cases in detail.

These forms of scarcity have different implications for benchmark design and are addressed in the design framework proposed in Section 5.

2.3 Research Questions

Building on past work, we organize this review around three research questions:

- **RQ1:** How do existing multilingual safety benchmarks perform across the dimensions of linguistic authenticity, cultural grounding, and evaluation transparency, as measured by a structured nine-dimension taxonomy?
- **RQ2:** Where do systematic gaps persist in the current benchmark landscape, and what structural forces produce them?
- **RQ3:** What design principles are necessary for safety evaluation to be valid in low-resource, non-Western linguistic contexts?

3 Methods: Taxonomy-Based Structured Review

3.1 Corpus Selection

Benchmarks were identified through systematic searches of major NLP and AI safety venues (ACL, EMNLP, NAACL, NeurIPS, ICLR, COLM, AAAI), preprint servers (arXiv, OpenReview), and government and industry technical reports. The primary focus was on work published between 2023 and 2025, reflecting rapid growth in multilingual safety evaluation. Earlier benchmarks were included where they establish methodological lineage relevant to the current landscape, including MLMA (Ousidhoum et al. 2019), AmericasNLI (Ebrahimi et al. 2022), and the Multilingual Content Moderation study (Ye et al. 2023). Inclusion required at least one of the following: a primary focus on safety, toxicity, or harm evaluation; multilingual or non-English scope; or English-only design with comparative relevance to multilingual safety evaluation, as in SALAD-Bench (Li et al. 2024) and WildGuard (Han et al. 2024), whose taxonomies and evaluation methodologies are widely adopted across multilingual work. Publication formats include peer-reviewed conference and journal papers as well as preprints.

The final corpus comprises 37 works. Of these, 27 are safety benchmarks designed to measure harmful, toxic, or unsafe model outputs, including hate speech detection, jailbreak resistance, toxicity scoring, and content moderation. These include broad evaluations (Ning et al. 2025; Li et al. 2024; Z. Zhang et al. 2024), targeted hate speech datasets (Muhammad et al. 2025; Ousidhoum et al. 2019), jailbreak evaluation frameworks (Yue Deng et al. 2024; Yoo, Yang, and Lee 2025; Liu et al. 2024), and deployed moderation tools (Kumar et al. 2025; Yihe Deng et al. 2026; Han et al. 2024). The remaining 10 are capability benchmarks evaluating general language understanding and reasoning without a primary safety focus, including IrokoBench (Adelani et al. 2025), AfriQA (Ogundepo et al. 2023), AfroBench (Ojo et al. 2025), BharatBench (Krutrim 2025), MILA (Sawarkar et al. 2026), MILU (Verma et al. 2025), AmericasNLI (Ebrahimi et al. 2022), PoETa v2 (Almeida et al. 2025), La Leaderboard (Grandury et al. 2025), and Cultural Expressiveness (LatAm) (Mora-Reyes, Drewyor, and Reyes-Angulo 2025).

Both categories are included because capability and safety benchmarks often draw on overlapping data sources and research communities in many Majority World contexts, and limitations in the underlying language ecosystem affect both. Capability benchmarks also demonstrate methodological practices less common in safety evaluation. AfriQA, for instance, draws on the same Masakhane research community as the safety benchmark AfriHate, but the two sit in largely separate literatures, while BharatBench mobilizes in-house linguists across 22 Indic languages at a scale no Indic safety benchmark has attempted. These works show that approaches to native data collection and community validation already exist in multilingual NLP. Capability benchmarks are therefore analyzed as methodological reference points rather than sources of safety-specific findings.

Beyond the safety and capability split, nine of the 37 works evaluate a single language rather than several. Two are the English-only benchmarks noted above (Li et al. 2024; Han et al. 2024), retained as methodological reference points. The other seven are single-language non-English benchmarks (Mubarak, Mohamed, and Hawasly 2025; Krasnodębska et al. 2025; Alam et al. 2026; Ridoy et al. 2026; Huang et al. 2024; W. Zhang et al. 2024; Almeida et al. 2025). These are included because the review examines how benchmarks represent language-specific harm and how they are constructed, and focused single-language efforts often demonstrate the native authorship, cultural adaptation, and community validation that broad multilingual benchmarks lack. Their single-language scope means they do not contribute to the cross-language comparison findings, which require multilingual benchmarks.

The corpus is therefore broader than the multilingual safety benchmarks named in the paper's central claim, and the cross-language safety findings in Section 4 are drawn only from the latter.

The corpus is predominantly text-based. Thirty-four of 37 works evaluate text modality only. The three exceptions are MuTox (Costa-jussà et al. 2024), which evaluates speech toxicity across 30 languages; Lingua-SafetyBench (Shi et al. 2026), which addresses

image-text multimodal safety across 10 languages; and BharatBench (Krutrim 2025), which spans text, vision, and speech for Indic languages. These examples highlight the limited development of multilingual safety evaluation outside text, despite Trust & Safety systems operating across multiple modalities in practice.

Language coverage ranges from single-language benchmarks for Arabic, Bengali, Chinese, and Polish to efforts covering 17 languages (Kumar et al. 2025), 28 languages (Wynter et al. 2025), 30 languages (Costa-jussà et al. 2024), and 64 African languages (Ojo et al. 2025). Data production methods fall into four categories: fully native-authored (8 works), fully translated from English (2 works), fully synthetic (2 works), and hybrid combinations of these (25 works). Table 2 presents the full taxonomy scoring and descriptive summary for all 37 works.

The corpus was assembled to cover multilingual and non-English safety evaluation broadly rather than to include only benchmarks that target low-resource languages. This is a deliberate design choice. Restricting the corpus to benchmarks that focus on low-resource languages would have measured the internal characteristics of that subset while obscuring the field-level pattern this review aims to document, namely how much of current multilingual safety evaluation reaches low-resource languages. Admitting the broader set of non-English and multilingual benchmarks allows low-resource coverage to be treated as a finding rather than an inclusion criterion, which makes it possible to report, in the following paragraphs, how few benchmarks center low-resource languages. The distinction matters because “multilingual” and “low-resource” are frequently conflated. A benchmark can cover many languages while still concentrating on high-resource ones, and only a corpus that admits both kinds can reveal that gap.

Mapping the corpus onto the three forms of scarcity defined in Section 2.2, 24 of the 37 works include at least one low-resource language, and 16 focus on low-resource languages as their primary target. The remaining 13 cover only high- and medium-resource languages. This coverage is uneven across the safety and capability split. Nine of the 10 capability benchmarks center low-resource languages, compared with 7 of the 27 safety benchmarks, indicating that native data collection for low-resource languages is more established in capability evaluation than in safety evaluation.

The three forms of scarcity overlap rather than partition the corpus, and a single benchmark often reflects more than one. Data scarcity maps most directly onto coverage. It applies to these 24 works (16 of them centrally), where limited digital text constrains data collection. Evaluation scarcity, the primary focus of this review, is a property of languages rather than individual works and cuts across the corpus, most visible in dedicated single-language efforts such as AraSafe (Mubarak, Mohamed, and Hawasly 2025), IndoSafety (Azmi et al. 2025), and BanglaGuard (Alam et al. 2026). Institutional scarcity is clearest in the multi-institutional and community-driven benchmarks, including the Masakhane-led African efforts (Adelani et al. 2025; Ojo et al. 2025) and RTP-LX (Wynter et al. 2025).

3.2 The Nine-Dimension Taxonomy

To enable structured comparison across benchmarks with different designs and goals, we developed a nine-dimension scoring taxonomy organized under three meta-dimensions. Each dimension is scored on a 0–2 scale, where 0 indicates absence or minimal engagement, 1 indicates partial or acknowledged engagement, and 2 indicates systematic engagement. The framework was developed iteratively. An initial version was applied to a subset of benchmarks, revised after reviewing ambiguous cases, and then applied to the full corpus.

The nine dimensions were derived from a broader set of attributes recorded during corpus review. Each benchmark was first coded on roughly two dozen descriptive fields spanning language coverage, data methodology, annotation practice, harm taxonomy, reporting, and access. These fields were then consolidated into the nine scored dimensions. Several descriptive fields were merged into a single dimension: data origin and translation method became Native Data; annotator profile and community involvement became Community Validation; code-switching handling and dialect coverage became Dialect and Code-Switch Coverage; and region-specific harms and taxonomy provenance became Local Harm Categories.

Other recorded attributes were retained as descriptive context rather than scored, because they characterize a benchmark without indicating the adequacy of its low-resource representation. Dataset size, the number of harm categories, and licensing, for instance, were recorded for every benchmark but not scored on the 0 to 2 scale. A larger dataset or a longer category list does not by itself indicate whether harmful language is natively expressed or whether harm categories are locally relevant. Native Data and Local Harm Categories capture those qualities directly.

The three meta-dimensions capture distinct aspects of benchmark design that this review treats as complementary rather than substitutable. Meta-Dimension A, Linguistic Authenticity, captures how faithfully a benchmark represents the target language as actually used, through three dimensions: Native Data Score, Dialect and Code-Switch Coverage, and Scarce Corpora Acknowledgment. Meta-Dimension B, Cultural Grounding, captures how well harm categories and validation processes reflect local context, through Cultural Adaptation, Local Harm Categories, and Community Validation. Meta-Dimension C, Evaluation Transparency and Operational Readiness, captures how usable the benchmark is for deployment decisions, through Disaggregated Metrics, Over-Refusal Analysis, and Operational Applicability. Table 1 provides the full scoring rubric.

The meta-dimensions are scored independently, with no weighting across them. A benchmark can score high on Linguistic Authenticity while scoring low on Evaluation Transparency, or vice versa. The taxonomy profiles benchmark strengths and weaknesses rather than producing a single ranking. Total scores range from a theoretical 0 to 18; in this corpus they span 3 (SALAD-Bench) to 14 (RabakBench, RTP-LX).

The taxonomy relates to existing safety frameworks by operating at a different level of analysis. Established harm taxonomies such as MLCommons and Llama Guard, and ethics benchmarks such as ETHICS and Scruples, specify what content counts as harmful and organize it into categories. This taxonomy does not classify harms and is not a substitute for these frameworks. It instead assesses how a benchmark is constructed: whether its data is natively authored, whether its harm categories are locally grounded or imported, and whether its results are reported in ways that reveal cross-language variation.

In this respect, the taxonomy is closer to dataset documentation practice (Bender and Friedman 2018; Gebru et al. 2021) than to harm classification, though it scores these properties comparatively across a corpus rather than documenting a single artifact. The relationship is made explicit in the Local Harm Categories dimension, which scores each benchmark by its stance toward these established taxonomies, from applying a Western taxonomy globally (0) to deriving categories locally (2). The framework therefore takes existing harm taxonomies as one of the objects it evaluates rather than as a standard it reproduces, which is what allows it to surface the cultural-grounding gaps documented in Section 4.

The nine dimensions are not claimed to be exhaustive. They were derived bottom-up from the attributes that recurred across the 37 benchmarks and consolidated from the broader set of coded fields described above, so they reflect the design choices that distinguish benchmarks in this corpus rather than a complete account of benchmark quality. The taxonomy is intended as a structured, repeatable review instrument rather than a certification standard, and the inter-rater results reported in Section 3.3 indicate how consistently it can be applied.

3.3 Scoring Procedure and Reliability

All 37 benchmarks were scored by the first author using the taxonomy described above. Scores were assigned through close reading of the main paper and, where available, supplementary materials and appendices. When methodological descriptions were ambiguous, scoring erred toward the lower value to avoid overstating benchmark rigor or coverage. Table 2 reports the full set of scores for all benchmarks in the corpus.

To assess the reliability of this single-evaluator procedure, a second author independently re-scored a stratified random subset of 12 of the 37 benchmarks (32%), selected to span the full range of total scores and to include both safety and capability benchmarks. The second author scored blind to the original ratings, applying the nine-dimension rubric to a standardized evidence summary compiled from each source paper.

Across the 98 doubly coded cells (excluding dimensions not applicable to capability benchmarks), exact agreement was 51% and agreement within one scale point was 92%; quadratic-weighted Cohen's κ (a chance-corrected measure of inter-rater agreement where 0 is chance-level and 1 is perfect) was 0.43, indicating moderate agreement.

Table 1: Scoring rubric for the nine evaluation dimensions. Dimensions are grouped under three meta-dimensions: Linguistic Authenticity, Cultural Grounding, and Evaluation Transparency.

Meta-Dimension	Dimension	0	1	2
A. Linguistic Authenticity	Native Data	Predominantly translated from another language	Hybrid (native plus translated or synthetic)	Substantially native-origin (authored, elicited, or native-language sourced)
	Dialect / Code-Switch Coverage	Standard variety only (a passing mention counts here)	Concrete but partial (a covered dialect/creole, code-mix tooling, or colloquial vs. formal)	Systematic (code-switching/ dialect as a core, varied condition)
	Scarce Corpora Acknowledgment	Assumes data availability	Acknowledges as limitation	Addresses with methodology
B. Cultural Grounding	Cultural Adaptation	Translated Western content, no adaptation	Post-hoc localization	Native design from the outset
	Local Harm Categories	Western taxonomy applied globally	Adapted Western with local additions	Locally derived categories
	Community Validation	No community involvement	Some native speaker review	Participatory design throughout
C. Evaluation Transparency	Disaggregated Metrics	Aggregate scores only	Some per-language results	Full per-language breakdown + analysis
	Over-Refusal Analysis	Not measured or discussed	Acknowledged or handled incidentally (no benign-input metric)	Explicitly measured (benign false-positive or oversensitivity metric)
	Operational Applicability	Academic only	Deployment-aware (evaluates deployed systems, ships a guard model, or concrete moderation framing)	Production artifact (deployed in service or integrated into a live pipeline)

Legend: 0 = absent/minimal, 1 = partial/acknowledged, 2 = systematic/comprehensive.

Agreement was highest for Local Harm Categories ($\kappa = 0.71$) and Scarce Corpora Acknowledgment ($\kappa = 0.67$) and lowest for Operational Applicability ($\kappa = 0.18$), Native Data ($\kappa = 0.21$), and Dialect/Code-Switching ($\kappa = 0.21$).

The 48 disagreements were adjudicated against the source papers; 30 upheld the original score and 18 revised it, with revisions concentrated in Operational Applicability and Over-Refusal Analysis. Because these disagreements stemmed largely from under-specified anchor boundaries rather than from differing readings of the papers, we refined the 0–2 definitions for the four lowest-agreement dimensions (Section 3.2) and applied the

refined definitions across the full corpus, updating the affected scores in Table 2. Broader limitations of the scoring approach and review design are discussed in Section 7.

4 Findings

Across the 37 works in the corpus, the taxonomy scores reveal uneven progress. Because Over-Refusal and Operational Applicability apply only to safety benchmarks, the pattern is clearest within the 27 safety benchmarks, with the 10 capability benchmarks summarized separately in Table 2. Scores are strongest on disaggregated reporting (1.44/2) and acknowledgment of data scarcity (1.33/2), and weakest on dialect and code-switching coverage (0.44/2), operational applicability (0.78/2), locally derived harm categories (0.81/2), and over-refusal analysis (0.93/2). Table 2 summarizes these scores for the full corpus. Sections 4.1–4.6 each examine one taxonomy dimension, addressing how benchmarks perform (RQ1) and where systematic gaps persist (RQ2); Section 4.7 synthesizes these patterns into recurring design archetypes that motivate the framework in Section 5 (RQ3). Each findings subsection closes with the design implication it supports. The subsections below unpack the patterns behind these averages.

4.1 Data Origins: Translation and Native Authorship

Translation from English is the dominant data production method across the corpus, including in benchmarks that present themselves as multilingual. MultiJail (Yue Deng et al. 2024) translates 315 English prompts into nine languages to construct a multilingual jailbreak dataset. CultureGuard (R. Joshi et al. 2025) produces 386K samples by culturally adapting English source data via Mixtral and then machine-translating the output. PolyGuard (Kumar et al. 2025) builds the largest multilingual safety training corpus to date at 1.91M samples, but 77% of that data is machine-translated from English-annotated sources. DuoGuard (Yihe Deng et al. 2026) claims support for 29 languages, yet its seed data is 81.4% English, with French, Spanish, and German collectively accounting for less than 19%.

Several studies document limitations of translation-based pipelines. SafetyBench (Z. Zhang et al. 2024) reports that ChatGPT refused to translate unsafe content and sometimes modified unsafe answers during translation, prompting the authors to use Baidu Translate instead. Even with a commercial API, the authors ultimately excluded the entire “Sensitive Topics” category because Chinese and English answers diverged systematically on political questions. These cases illustrate how translation pipelines can modify the content they are intended to preserve and how safety categories developed in one context may not transfer cleanly to another.

IndoSafety (Azmi et al. 2025) reaches a similar conclusion from the opposite direction. The authors argue that translated safety prompts overlook informal registers and

Table 2: Heatmap scoring matrix evaluating all 37 works in the corpus across the nine dimensions defined in Table 1, organized under three meta-dimensions. The taxonomy is designed to profile benchmark strengths and weaknesses rather than produce a single ranking. *Capability benchmarks* are scored on applicable dimensions only; Over-Refusal and Operational Applicability do not apply and are marked "—". Subgroup averages are reported separately for the 27 safety and 10 capability benchmarks.

	Linguistic Authenticity			*Cultural Grounding*			*Evaluation Transparency*			
Benchmark	Native Data	Dialect/Code-Switch	Scarce Corpora	Cultural Adapt.	Local Harm Cat.	Community Valid.	Disaggregation	Over-Refusal	Operational	**Total**
RabakBench	1	2	2	2	2	1	2	1	1	**14**
RTP-LX	1	1	2	2	1	2	2	2	1	**14**
AfriHate	2	1	2	2	2	1	2	0	1	**13**
SGToxicGuard	2	2	1	2	2	1	2	0	0	**12**
IndoSafety	1	1	2	2	2	1	2	1	0	**12**
LinguaSafe	1	0	2	1	1	1	2	2	1	**11**
BengaliMoralBench	2	0	2	2	2	2	0	0	1	**11**
La Leaderboard	1	1	2	1	1	2	2	—	—	**10**
AfriQA	1	0	2	2	0	2	2	—	—	**9**
HASOC (2023–2024)	2	0	2	1	1	1	2	0	0	**9**
BanglaGuard	1	0	2	1	1	1	0	2	1	**9**
MLMA	2	1	1	1	1	1	2	0	0	**9**
CSRT	0	2	2	0	0	0	2	2	1	**9**
Chai Safety Bench	1	0	0	2	2	0	2	1	1	**9**
AraSafe	1	1	2	2	1	1	0	0	0	**8**
S-Eval	1	0	1	1	1	0	1	1	2	**8**
FLAMES	2	0	0	2	2	1	0	1	0	**8**
MultiJail	1	0	2	0	0	1	2	1	1	**8**
PolyGuard	1	1	1	0	0	1	2	1	1	**8**
Cultural Expressiveness (LatAm)	2	0	2	1	1	1	0	—	—	**7**
PoETa v2	1	1	1	1	0	1	2	—	—	**7**
BharatBench	1	0	1	2	0	1	2	—	—	**7**
MILA	1	0	2	1	0	1	2	—	—	**7**
AfroBench	1	0	2	1	0	1	2	—	—	**7**
CultureGuard	0	0	2	1	0	0	2	1	1	**7**
MILU	1	0	1	2	0	1	2	—	—	**7**
Lingua-SafetyBench	1	0	1	0	0	1	2	1	1	**7**
PL-Guard	1	0	2	1	0	1	0	1	1	**7**
Multilingual Mod. (Reddit)	2	0	1	0	0	0	2	1	1	**7**
WildGuard	1	0	0	0	1	1	0	2	1	**6**
AmericasNLI	0	0	2	1	0	1	2	—	—	**6**
IrokoBench	0	0	2	1	0	1	2	—	—	**6**
MuTox	1	0	1	0	0	1	2	0	1	**6**
DuoGuard	0	0	2	0	0	0	2	1	1	**6**
JailJudge	0	0	1	0	0	0	2	1	1	**5**
SafetyBench	1	0	0	1	0	0	2	1	0	**5**
SALAD-Bench	0	0	0	0	0	1	0	1	1	**3**
Safety Benchmarks avg. (n=27)	1.07	0.44	1.33	0.96	0.81	0.78	1.44	0.93	0.78	
Capability Benchmarks avg. (n=10)	0.90	0.20	1.70	1.30	0.25	1.20	1.80	—	—	

Legend: 0 = absent/minimal, 1 = partial/acknowledged, 2 = systematic/comprehensive; *italic* = capability (non-safety) benchmark; "—" = not applicable.

culturally specific norms. Their evaluation shows that models appearing safe on formal Indonesian prompts produce unsafe responses at substantially higher rates when prompted in colloquial Indonesian, Javanese, Sundanese, or Minangkabau.

The most sophisticated attempt to address translation distortion in the corpus is LinguaSafe's TATER (Task-Aware Translate, Estimate and Refine) framework (Ning et al. 2025), which combines machine translation with LLM-based transcreation and human refinement. TATER reduces Bengali translation error rates from 71% to 12% and Malay error rates from 36% to 3%. However, the benchmark's harm taxonomy remains Western-derived and the data pipeline still begins with English source material.

Several benchmarks demonstrate that native-authored data collection is feasible and produces qualitatively different results. AraSafe (Mubarak, Mohamed, and Hawasly 2025) collects 12K prompts from 110 native Arabic speakers across eight countries. AfriHate (Muhammad et al. 2025) assembles approximately 75K items across 15 African languages using culture-specific keywords and native annotators. BengaliMoralBench (Ridoy et al. 2026) takes this furthest, with 3,000 items handcrafted by 30 Bengali speakers covering culturally grounded scenarios such as load-shedding etiquette and the distinction between zakat and voluntary charity. These scenarios cannot easily be reproduced through translation of English-language ethics benchmarks such as ETHICS or Scruples. Across these works, native-authored benchmarks consistently surface vulnerabilities that translated benchmarks miss (Ning et al. 2025; Azmi et al. 2025).

The countervailing tension is that native data collection is expensive and difficult to scale. AfriHate required collaboration across more than 20 institutions funded by the Lacuna Fund (Muhammad et al. 2025), while RTP-LX (Wynter et al. 2025) paid professional transcreators between 19 and 54 USD per hour across 28 languages.

Native authorship also does not guarantee benchmark quality. AraSafe's participant pool skews toward Egypt, with roughly two-thirds of contributors from a single country (Mubarak, Mohamed, and Hawasly 2025). AfriHate reports class imbalance in several languages, with the hate speech class containing fewer than 200 instances in some cases (Muhammad et al. 2025). Issues of geographic representation, annotation consistency, and taxonomy design remain relevant even in native-authored datasets.

This tension reflects a central design trade-off in multilingual safety evaluation. Translation pipelines scale easily but can distort culturally embedded harms. Native authorship produces more contextually grounded data but is slower and more resource-intensive. Effective benchmark design therefore requires combining both approaches while reporting their trade-offs transparently.

4.2 Aggregate Metrics and Hidden Cross-Language Variation

Even benchmarks that report per-language results often fail to analyze the disparities those results reveal. Performance differences across languages can be substantial. RabakBench (Chua et al. 2026) finds that WildGuard achieves 78.89% F1 (the harmonic mean of precision and recall, where higher values indicate more accurate harmful-content detection) on Singlish but drops to 2.32% on Tamil, despite both languages being used by the same population on the same platform. AWS Bedrock Guardrails shows a similar collapse, from 66.50% F1 on Singlish to 0.06% on Chinese (Chua et al. 2026). LinguaSafe (Ning et al. 2025) reveals a more complex pattern. Claude shows lower vulnerability in Arabic and Thai than in English, but higher oversensitivity in those same languages, indicating misalignment across error types rather than a simple safety gradient.

These disparities become invisible when results are reported only in aggregate. SALAD-Bench (Li et al. 2024) and WildGuard (Han et al. 2024) are English-only, while DuoGuard (Yihe Deng et al. 2026) evaluates only four high-resource languages despite its base model supporting 29. Aggregate metrics can therefore mask meaningful differences in model behavior across linguistic contexts.

The relationship between language resource level and safety performance is not strictly deterministic. The HASOC shared task analysis (Ghosh et al. 2026) provides an instructive example. Across four languages (Assamese, Bengali, Bodo, and English), Bodo, the lowest-resource language in the set, achieved the highest best-system F1 at 0.85, compared to 0.81 for English. The authors attribute this partly to dataset construction quality. The Bodo dataset had the highest inter-annotator agreement (Fleiss's $\kappa = 0.633$, compared to 0.366 for English) and the lowest average item difficulty. This suggests that evaluation design and annotation quality can partially mitigate resource disparities, even if they do not eliminate the underlying data gap.

These patterns underscore the importance of disaggregated reporting. When cross-lingual variation is substantial, aggregate metrics provide an incomplete picture of system performance. Per-language breakdowns, accompanied by analysis of where and why performance gaps emerge, are therefore necessary for benchmarks intended to inform multilingual safety decisions.

4.3 Imported Harm Taxonomies and Cultural Misalignment

The dominant safety frameworks in the corpus originate in Western policy environments and are applied globally with minimal adaptation. PolyGuard (Kumar et al. 2025) applies the MLCommons Safety Taxonomy across 17 languages without cultural validation. CultureGuard (R. Joshi et al. 2025) describes its approach as "cultural adaptation," but in practice this involves substituting institutional references (such as replacing SSN with Aadhaar) while retaining the underlying Western harm ontology.

As a result, certain forms of harm remain unmeasured. IndoSafety (Azmi et al. 2025)

identifies seven Indonesia-specific safety categories, including misinterpretation of Pancasila (the national philosophical foundation), regional separatism, and supernatural beliefs, none of which appear in MLCommons or Llama Guard taxonomies. AfriHate (Muhammad et al. 2025) documents hate speech patterns tied to regional political targeting and disability-based discrimination in the political contexts of 15 African countries. BengaliMoralBench (Ridoy et al. 2026) constructs a culturally grounded ethical framework rooted in Bangladeshi social norms that cannot be directly mapped onto Western benchmarks such as ETHICS or Scruples.

FLAMES (Huang et al. 2024) illustrates the practical implications of this gap. The benchmark integrates Chinese philosophical values, including harmony (和谐), benevolence (仁), and the Doctrine of the Mean (中庸), alongside standard safety dimensions. When tested on ChatGPT, FLAMES achieves a 53% attack success rate (the proportion of adversarial prompts that elicit an unsafe or policy-violating response, where higher values indicate weaker safety), compared to 1.6% for Safety-prompts and 3.1% for CValues, two Chinese-language benchmarks that rely on simpler prompt designs without culturally grounded adversarial framing. These results suggest that models performing well on translated safety prompts may still fail when evaluated against scenarios requiring knowledge of local values and social norms.

This pattern does not imply that harm itself is culturally relative. However, the ways harms are expressed, through language, social references, or political sensitivities, vary across communities; Jiang et al. (2021) find that perceptions of harm severity differ systematically across countries. A taxonomy built for one context will not capture these variations in another. When evaluation frameworks rely primarily on Western-derived categories, gaps in coverage tend to fall on the linguistic communities least represented in their design.

4.4 Limited Attention to Over-Refusal and False Positives

Safety evaluation has two directions of error. Systems may fail to flag harmful content or incorrectly classify benign content as harmful. The latter suppresses legitimate speech and degrades user experience, the free-expression side of the central content-moderation trade-off between removing harm and preserving legitimate speech (Kozyreva et al. 2023). Despite this, over-refusal is rarely measured in the multilingual safety evaluation corpus. Of the 27 safety benchmarks reviewed, 7 score zero on the Over-Refusal Analysis dimension, meaning they did not assess false positives. The safety-benchmark average on this dimension is 0.93 out of 2.

Only five benchmarks systematically measure over-refusal. LinguaSafe (Ning et al. 2025) is the most comprehensive, incorporating oversensitivity evaluation alongside vulnerability assessment and showing that Claude exhibits lower vulnerability in Arabic and Thai than in English but higher oversensitivity in those languages. BanglaGuard (Alam et al. 2026) tracks refusal rate and unsafe completion rate (the share of benign prompts

the model declines and the share of harmful prompts it answers) as separate metrics. RTP-LX (Wynter et al. 2025) reports false positive rates across evaluator models, finding that Gemma 2B misidentifies up to 40% of benign completions as toxic while Llama Guard produces near-zero false positives. WildGuard (Han et al. 2024) includes refusal detection as a distinct task alongside harm detection, though its evaluation is English-only.

CSRT (Yoo, Yang, and Lee 2025) is the fifth, quantifying benign over-filtering directly. Perplexity-based defense filters, which flag inputs whose statistical unfamiliarity to the model is high on the assumption that adversarial text is less predictable, incorrectly block 87.1% of benign Bengali queries and 97.9% of benign code-switching queries, because unfamiliar linguistic patterns generate high perplexity scores that the filter interprets as adversarial.

For Trust & Safety teams, it can be challenging to distinguish between genuinely robust systems and those that suppress benign speech in unfamiliar linguistic contexts. In low-resource languages, where training data is sparse and linguistic patterns diverge most from the English-dominated training distribution, over-refusal may be a dominant failure mode. Limited attention to this dimension means that the communities most affected by such behavior are also those least represented in current evaluation frameworks.

4.5 Dialect, Code-Switching, and Informal Language Coverage

Real language use in Majority World contexts routinely involves code-switching and dialect variation. Singlish weaves English, Malay, Hokkien, and Tamil within single sentences. Hindi-English code-switching is a common register for millions of South Asian internet users. Arabic-speaking communities operate across Modern Standard Arabic and numerous regional dialects that differ substantially in vocabulary and syntax. Despite this, the multilingual safety evaluation corpus primarily treats “standard” written varieties as adequate proxies for everyday communication. Of the 27 safety benchmarks, 18 score zero on the Dialect and Code-Switching dimension, meaning they make no accommodation for linguistic variation beyond standard written forms. The safety-benchmark average is 0.44 out of 2, the lowest dimension in the taxonomy.

Only three works score a 2 on this dimension. SGToxicGuard (Hu et al. 2025) and RabakBench (Chua et al. 2026) are both built for Singapore, where Singlish functions as a recognized creole that cannot be meaningfully evaluated by treating English, Malay, Tamil, and Chinese as separate monolingual streams. These benchmarks treat code-mixing as a central design requirement rather than an edge case. The third, CSRT (Yoo, Yang, and Lee 2025), approaches code-switching from the perspective of jailbreak attacks. CSRT shows that intra-sentence code-switching increases attack success rates by 46.7% over monolingual English attacks, with success increasing as more languages are mixed in a single query. Lower-resource languages in the mix correlate with higher attack success. Although framed as a jailbreak vulnerability, the finding also suggests that safety

alignment weakens in the linguistic environments where multilingual users commonly communicate.

Several other works engage with linguistic variation partially. IndoSafety (Azmi et al. 2025) evaluates formal versus colloquial Indonesian alongside Javanese, Sundanese, and Minangkabau, finding that models perform substantially worse on colloquial and local-language prompts. AfriHate (Muhammad et al. 2025) builds a custom language identification model to handle code-mixing, though the benchmark does not systematically vary code-switching as an evaluation condition.

The strongest work on this dimension is geographically concentrated. Both SGToxicGuard and RabakBench emerge from Singapore, where Singlish's status as a recognized national vernacular and government investment in AI safety have supported benchmarks that treat code-mixing as a core design requirement. Comparable safety benchmarks for other regions where code-switching is equally pervasive, including South Asia, West Africa, and broader Southeast Asia, do not appear in the selected corpus.

4.6 Operational and Deployment-Relevant Evaluation

Trust & Safety systems operate under conditions that academic benchmarks rarely reproduce: heavily imbalanced class distributions, adversarial inputs, and latency or cost constraints. The Operational Applicability dimension captures how far a benchmark engages these conditions. Because it does not apply to capability benchmarks, it is scored only for the 27 safety benchmarks, where it scores low (average 0.78/2).

Only one benchmark, S-Eval (Yuan et al. 2025), reports validation in an actual production setting, having been deployed by an industrial partner for automated safety evaluation at scale. A second group (19 of 27) engages deployment at one remove. Some evaluate already-deployed moderation systems, such as RabakBench (Chua et al. 2026), which tests 13 commercial and open-source guardrails under adversarial and code-mixed inputs, and CSRT (Yoo, Yang, and Lee 2025). Others release a guard model that could itself be deployed, including WildGuard (Han et al. 2024), PolyGuard (Kumar et al. 2025), and CultureGuard (R. Joshi et al. 2025). The remaining 7 are academic resources with no deployment evaluation.

The benchmarks strongest on cultural grounding sit largely in this academic group. BengaliMoralBench (Ridoy et al. 2026), AraSafe (Mubarak, Mohamed, and Hawasly 2025), and AfriHate (Muhammad et al. 2025) score 0 or 1 on this dimension, reflecting an orientation toward contextual validity rather than deployment workflows. The corpus therefore provides little evidence about how any benchmark, least of all the most culturally grounded ones, behaves in a live moderation pipeline, where decisions depend on platform rules, context, and locally varying norms rather than on balanced, decontextualized prompts (Ye et al. 2023). Moderation is better assessed as a system than as a series of individual decisions (Douek 2022), and it is at the system level that the corpus offers

least evidence. This is the central tension synthesized in the next subsection.

4.7 Synthesis: Benchmark Archetypes and Design Trade-Offs

The preceding subsections examined the corpus one dimension at a time. Read together, they show that benchmarks cluster into a small number of design strategies, each making a coherent trade-off across dimensions rather than excelling uniformly. The data reveals three such archetypes, each representing a genuine trade-off rather than a single best practice.

The first archetype is the participatory native benchmark, exemplified by AfriHate (Muhammad et al. 2025), AraSafe (Mubarak, Mohamed, and Hawasly 2025), BengaliMoralBench (Ridoy et al. 2026), and SGToxicGuard (Hu et al. 2025). These benchmarks score highest on Linguistic Authenticity and Cultural Grounding. AfriHate and SGToxicGuard each score 5 out of 6 on both meta-dimensions, while BengaliMoralBench achieves the highest Cultural Grounding score in the corpus (6 out of 6) due to its fully native authorship, participatory design process, and locally derived ethical framework. The trade-off appears in Evaluation Transparency. BengaliMoralBench scores only 1 on that meta-dimension (reporting no disaggregated metrics or over-refusal analysis, and engaging deployment only through the binary framing of its scenarios) while AraSafe scores 0. These benchmarks produce highly contextualized data but tend to be narrow in scope, expensive to build, and less connected to deployment workflows.

The second archetype is the transcreated hybrid benchmark, represented by RTP-LX (Wynter et al. 2025), RabakBench (Chua et al. 2026), IndoSafety (Azmi et al. 2025), and LinguaSafe (Ning et al. 2025). These approaches combine native authorship or professional transcreation with translated or synthetic data and invest in methodological rigor across several dimensions. RabakBench and RTP-LX both score 14, reflecting strong performance across all three meta-dimensions. This archetype provides the clearest balance between cultural validity and evaluation infrastructure. The trade-off is cost and complexity. RTP-LX required professional transcreators across 28 languages, RabakBench involved multi-stage human calibration of LLM labelers, and LinguaSafe developed the TATER framework to improve transcreation quality. Such approaches require substantial institutional investment that may be difficult to access in low-resource settings.

The third archetype is the synthetic adaptation pipeline, exemplified by CultureGuard (R. Joshi et al. 2025), DuoGuard (Yihe Deng et al. 2026), and Lingua-SafetyBench (Shi et al. 2026). These approaches prioritize scale and automation, producing large multilingual datasets through synthetic generation, machine translation, or both. CultureGuard produces 386K samples across nine languages, while DuoGuard generates synthetic training data through a two-player reinforcement learning framework. These benchmarks score relatively well on Evaluation Transparency (4 out of 6) but poorly on Cultural Grounding. CultureGuard scores 1, DuoGuard scores 0, and Lingua-SafetyBench scores

1. CultureGuard does demonstrate a 31% improvement over English-only safety models (R. Joshi et al. 2025), suggesting that synthetic pipelines can expand language coverage even if they rarely surface culturally specific harm categories or dialect variation.

Taken together, the results suggest that no single strategy performs well across all dimensions. Effective multilingual safety evaluation therefore depends on combining these approaches while reporting their trade-offs transparently. These trade-offs are not incidental; they are produced by the structural forces that shape the field (RQ2): the cost of native data collection, the institutional investment required for professional transcreation, and the automation economics of synthetic generation. The design framework in Section 5 translates each dimension-level finding into a corresponding principle, while recognizing that no single benchmark is likely to maximize all of them at once.

5 Design Framework for Multilingual Safety Benchmarks

The patterns identified in Section 4 highlight a distinction between linguistic coverage and cultural grounding. Coverage alone does not indicate whether evaluation data captures how harm is expressed in different linguistic communities or whether harm categories reflect locally relevant concerns. This gap between what a benchmark measures and what it is taken to represent is a problem of construct validity (Raji et al. 2021).

The following six principles draw on the empirical patterns identified in Section 4 and provide practical guidance for multilingual safety benchmark design. Each principle corresponds directly to a findings subsection in Section 4, so that every design recommendation is traceable to an observed pattern in the corpus.

These principles also respond to the three forms of scarcity distinguished in Section 2.2, each of which carries a different design implication. Data scarcity, the limited availability of digital and training text in a language, is the condition that translation-based pipelines and standard-variety evaluation most often obscure, and it motivates the native-authored corpus requirement (Section 5.1) and the treatment of dialect and code-switching as core evaluation conditions (Section 5.5).

Evaluation scarcity, the absence of purpose-built benchmarks for a language, is what aggregate reporting conceals and what unmeasured over-refusal leaves undetected, and it motivates disaggregated reporting (Section 5.2) and the measurement of false positives (Section 5.4). Institutional scarcity, the limited annotation capacity and community representation available for a language, constrains both the derivation of locally grounded harm categories (Section 5.3) and the deployment testing that operational validity requires (Section 5.6). Each principle therefore targets a specific form of resource limitation identified in Section 2.2.

5.1 Native-Authored Minimum Viable Corpus

Benchmarks claiming multilingual coverage should include a minimum threshold of natively authored data in the target language. Hybrid approaches combining native, translated, or synthetic data are often necessary for scale, but the relative contribution of each source should be reported transparently. Evidence across the corpus suggests that translation-only pipelines frequently miss culturally embedded harm expression (Ning et al. 2025). Native-authored data provides an important foundation for evaluating harms embedded in everyday language use.

5.2 Disaggregated Reporting as Default

Benchmarks should report per-language performance as a default publication standard rather than as supplementary material. Aggregate metrics can obscure substantial performance differences across languages (Chua et al. 2026). Disaggregated evaluation is already established practice in model reporting (Mitchell et al. 2019); what it requires here is disaggregation by language. Per-language analysis improves transparency for both researchers and practitioners selecting evaluation tools.

5.3 Locally Grounded Harm Taxonomies

Benchmark harm categories should be derived from, or validated with, communities familiar with the linguistic and cultural context of the target language. Established taxonomies can provide useful starting points, but they often require adaptation to capture locally relevant harms. IndoSafety (Azmi et al. 2025), for example, introduces categories related to Indonesian political and cultural contexts that do not appear in widely used Western safety taxonomies. Community involvement during dataset design can therefore improve the cultural relevance of benchmark content and help ensure that evaluation reflects the environments in which models are deployed.

5.4 Measuring False Positives

Safety evaluation should measure both harmful outputs and over-refusal rates. Much of the existing evaluation literature focuses primarily on harmful completions, leaving false positives comparatively underexamined. Measuring both error directions provides a more complete view of system behavior.

5.5 Dialect and Code-Switching Coverage

Benchmarks should treat dialectal variation and code-switching as core evaluation conditions rather than edge cases, particularly for languages whose speakers routinely mix codes in everyday communication. Evaluations restricted to standard written varieties can overstate safety. CSRT (Yoo, Yang, and Lee 2025) finds that intra-sentence code-switching increases attack success rates by 46.7% over monolingual English attacks,

and the guardrail collapse across languages reported in Section 4.2 shows the same pattern in deployed systems. Where feasible, benchmarks should include code-mixed and colloquial inputs and report performance on them separately, as SGToxicGuard (Hu et al. 2025) and RabakBench do by design.

5.6 Operational Validity Testing

Building on the operational patterns documented in Section 4.6, benchmarks intended to inform Trust & Safety deployment should evaluate models under conditions that approximate real platform environments. Academic datasets often rely on balanced class distributions and controlled prompts, while production moderation systems operate under highly imbalanced conditions and adversarial inputs. RabakBench (Chua et al. 2026) illustrates this difference by evaluating guardrail systems using multilingual content and adversarial prompts that better reflect platform moderation contexts. Evaluation under deployment-relevant conditions complements traditional benchmark design.

6 Implications for Trust & Safety Practice

The findings of this review have practical implications for Trust & Safety practitioners, including platform policy teams, guardrail model designers, frontier labs, and regulators.

Benchmark selection carries governance implications, since automated classifiers and guardrails are themselves instruments of platform governance rather than neutral technical tools (Gorwa, Binns, and Katzenbach 2020; Gillespie 2018). Platforms serving multilingual user bases that rely on English-origin evaluation tools implicitly treat English-language performance as a proxy for safety across other languages. Documenting evaluation coverage explicitly can help organizations understand the limits of their safety assessments.

Evaluation coverage also shapes the ability to detect harm. When evaluation tools do not include certain categories of harm, it becomes difficult to assess whether deployed systems are robust against those risks. As documented in Section 4, several benchmarks identify vulnerabilities invisible to English-derived evaluation frameworks. Reporting evaluation coverage is therefore as important as reporting benchmark performance.

The most culturally grounded benchmarks in this corpus emerged from collaborations involving regional research communities and targeted funding. The expertise to design locally grounded benchmarks is widely distributed. The labor of commercial content moderation is itself globally distributed (Roberts 2019), often to regions whose languages remain least represented in evaluation. What is often limited is the institutional infrastructure to support evaluation development at scale.

For frontier labs, partnerships with researchers and institutions working in low-resource language environments can strengthen the empirical foundations of safety evaluation. For platform operators, extending evaluation coverage to the languages represented in their user bases can improve the reliability of system-level safety assessments. For regulators considering AI safety standards, evaluation requirements that account for variation across language resource levels may better reflect the conditions under which systems are deployed globally.

7 Limitations

The findings and framework presented in this review should be interpreted in light of several methodological and scope limitations.

First, the scoring procedure relied on a single primary evaluator. All 37 benchmarks were initially scored by the first author. To assess reliability, a second author independently re-scored a 32% subset blind to the original ratings (Section 3.3); inter-rater agreement was moderate (quadratic-weighted $\kappa = 0.43$; 92% within one scale point), the 48 disagreements were adjudicated against the source papers, and the lowest-agreement dimensions were given sharper definitions that were re-applied across the corpus. Nonetheless, agreement before adjudication was only moderate, full-corpus multi-rater scoring was not performed, and the 0–2 scale leaves residual interpretive latitude; the scores should therefore be read as a structured and partially validated characterization rather than a definitive measurement.

Second, the review corpus is weighted toward benchmarks published in English-language venues and widely indexed conference proceedings. Work published primarily in other languages or regional venues may therefore be underrepresented. In addition, the benchmark landscape is evolving rapidly, and several works included in the review were still in preprint at the time of scoring. The results should therefore be interpreted as a snapshot of the field as it existed in late 2025.

Third, the nine-dimension taxonomy reflects the evaluative priorities of this review and therefore represents an interpretive framework rather than a universal standard for benchmark quality. The 0–2 scoring scale also compresses variation within each dimension, meaning that benchmarks with different methodological choices may receive the same score.

Finally, the design framework proposed in Section 5 is derived from patterns observed in the benchmark corpus rather than from demonstrated causal relationships between benchmark design and downstream moderation performance. The extent to which improvements in benchmark design translate into measurable improvements in deployed safety systems remains an open empirical question.

8 Conclusion

This paper reviewed 37 benchmarks across safety and capability evaluation using a nine-dimension taxonomy spanning linguistic authenticity, cultural grounding, and evaluation transparency. Multilingual coverage has expanded, but the corpus shows limited attention to dialect and code-switching coverage, operational readiness, over-refusal analysis, and locally derived harm categories. Representation in evaluation datasets does not ensure that benchmarks capture how harm is expressed across linguistic and cultural contexts.

Several benchmarks in the corpus, including RabakBench (Chua et al. 2026), RTP-LX (Wynter et al. 2025), AfriHate (Muhammad et al. 2025), and IndoSafety (Azmi et al. 2025), demonstrate that culturally grounded multilingual evaluation is achievable.

The six-principle design framework proposed in Section 5 builds on these examples, outlining directions for strengthening multilingual safety evaluation as LLMs continue to be deployed across diverse linguistic environments.

References

Adelani, David Ifeoluwa, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, et al. 2025. "IrokoBench: A New Benchmark for African Languages in the Age of Large Language Models." In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers),* 2732–57. https://doi.org/10.18653/v1/2025.naacl-long.139.

Alam, Md Jahangir, Tanzim Ahad, James Newson, Ismail Hossain, Sai Puppala, and Sajedul Talukder. 2026. "BanglaGuard: Benchmarking and Defending Large Language Models for Safety in Low-Resource Languages." Preprint. Accessed August 7, 2026. https://openreview.net/forum?id=KTsGJzaEPg.

Almeida, Thales Sales, Ramon Pires, Hugo Abonizio, Rodrigo Nogueira, and Hélio Pedrini. 2025. "PoETa v2: Toward More Robust Evaluation of Large Language Models in Portuguese." *arXiv,* https://doi.org/10.48550/arXiv.2511.17808.

Azmi, Muhammad Falensi, Muhammad Dehan Al Kautsar, Alfan Farizki Wicaksono, and Fajri Koto. 2025. "IndoSafety: Culturally Grounded Safety for LLMs in Indonesian Languages." In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing,* 9146–77. https://doi.org/10.18653/v1/2025.emnlp-main.465.

Bender, Emily M., and Batya Friedman. 2018. "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science." *Transactions of the Association for Computational Linguistics* 6:587–604. https://doi.org/10.1162/tacl_a_00041.

Chua, Gabriel, Leanne Tan, Ziyu Ge, and Roy Ka-Wei Lee. 2026. "Lost in Localization: Building RabakBench with Human-in-the-Loop Validation to Measure Multilingual Safety Gaps." *arXiv,* https://doi.org/10.48550/arXiv.2507.05980.

Costa-jussà, Marta R., Mariano Coria Meglioli, Pierre Andrews, David Dale, Prangthip Hansanti, Elahe Kalbassi, Alex Mourachko, Christophe Ropers, and Carleigh Wood. 2024. "MuTox: Universal Multilingual Audio-Based Toxicity Dataset and Zero-Shot Detector." *arXiv,* https://doi.org/10.48550/arXiv.2401.05060.

Deng, Yihe, Yu Yang, Junkai Zhang, Wei Wang, and Bo Li. 2026. "Enhancing LLM Safety through a Theoretical Minimax Game Lens." *arXiv,* https://doi.org/10.48550/arXiv.2502.05163.

Deng, Yue, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2024. "Multilingual Jailbreak Challenges in Large Language Models." *arXiv,* https://doi.org/10.48550/arXiv.2310.06474.

Douek, Evelyn. 2022. "Content Moderation as Systems Thinking." *Harvard Law Review* 136 (2): 526–607. https://www.jstor.org/stable/27339278.

Ebrahimi, Abteen, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, et al. 2022. "AmericasNLI: Evaluating Zero-Shot Natural Language Understanding of Pretrained Multilingual Models in Truly Low-Resource Languages." In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 6279–99. https://doi.org/10.18653/v1/2022.acl-long.435.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92. https://doi.org/10.1145/3458723.

Ghosh, Koyel, Saptarshi Saha, Thomas Mandl, and Sandip Modha. 2026. "Findings from Shared Tasks on Hate Speech Detection: Performance Patterns for Low-Resource Languages." *Pattern Recognition Letters* 199:303–9. https://doi.org/10.1016/j.patrec.2025.09.004.

Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media.* Yale University Press. https://doi.org/10.12987/9780300235029.

Gorwa, Robert, Reuben Binns, and Christian Katzenbach. 2020. "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance." *Big Data & Society* 7 (1). https://doi.org/10.1177/2053951719897945.

Grandury, María, Javier Aula-Blasco, Júlia Falcão, Clémentine Fourrier, Miguel González Saiz, Gonzalo Martínez, Gonzalo Santamaria Gomez, et al. 2025. "La Leaderboard: A Large Language Model Leaderboard for Spanish Varieties and Languages of Spain and Latin America." In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 32482–524. https://doi.org/10.18653/v1/2025.acl-long.1561.

Han, Seungju, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. 2024. "WildGuard: Open One-Stop Moderation Tools for Safety Risks, Jailbreaks, and Refusals of LLMs." *arXiv,* https://doi.org/10.48550/arXiv.2406.18495.

Hu, Yujia, Ming Shan Hee, Preslav Nakov, and Roy Ka-Wei Lee. 2025. "Toxicity Red-Teaming: Benchmarking LLM Safety in Singapore's Low-Resource Languages." In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing,* 12194–212. https://doi.org/10.18653/v1/2025.emnlp-main.612.

Huang, Kexin, Xiangyang Liu, Qianyu Guo, Tianxiang Sun, Jiawei Sun, Yaru Wang, Zeyang Zhou, et al. 2024. "Flames: Benchmarking Value Alignment of LLMs in Chinese." *arXiv,* https://doi.org/10.48550/arXiv.2311.06899.

Jiang, Jialun Aaron, Morgan Klaus Scheuerman, Casey Fiesler, and Jed R. Brubaker. 2021. "Understanding International Perceptions of the Severity of Harmful Content Online." *PLOS ONE* 16 (8). https://doi.org/10.1371/journal.pone.0256762.

Joshi, Pratik, 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,* 6282–93. https://doi.org/10.18653/v1/2020.acl-main.560.

Joshi, Raviraj, Rakesh Paul, Kanishk Singla, Anusha Kamath, Michael Evans, Katherine Luna, Shaona Ghosh, et al. 2025. "CultureGuard: Towards Culturally-Aware Dataset and Guard Model for Multilingual Safety Applications." *arXiv,* https://doi.org/10.48550/arXiv.2508.01710.

Kozyreva, Anastasia, Stefan M. Herzog, Stephan Lewandowsky, Ralph Hertwig, Philipp Lorenz-Spreen, Mark Leiser, and Jason Reifler. 2023. "Resolving Content Moderation Dilemmas between Free Speech and Harmful Misinformation." *Proceedings of the National Academy of Sciences* 120 (7). https://doi.org/10.1073/pnas.2210666120.

Krasnodębska, Aleksandra, Karolina Seweryn, Szymon Łukasik, and Wojciech Kusa. 2025. "PL-Guard: Benchmarking Language Model Safety for Polish." In *Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025),* 25–37. https://doi.org/10.18653/v1/2025.bsnlp-1.4.

Krutrim AI Team. 2025. *BharatBench: Comprehensive Multilingual Multimodal Evaluations of Foundation AI Models for Indian Languages.* Krutrim AI Labs. https://ai-labs.olakrutrim.com/static/Bharatbench-report-4thfeb.pdf.

Kumar, Priyanshu, Devansh Jain, Akhila Yerukola, Liwei Jiang, Himanshu Beniwal, Thomas Hartvigsen, and Maarten Sap. 2025. "PolyGuard: A Multilingual Safety Moderation Tool for 17 Languages." *arXiv,* https://doi.org/10.48550/arXiv.2504.04377.

Li, Lijun, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. 2024. "SALAD-Bench: A Hierarchical and Comprehensive Safety Benchmark for Large Language Models." *arXiv,* https://doi.org/10.48550/arXiv.2402.05044.

Liu, Fan, Yue Feng, Zhao Xu, Lixin Su, Xinyu Ma, Dawei Yin, and Hao Liu. 2024. "JailJudge: A Comprehensive Jailbreak Judge Benchmark with Multi-Agent Enhanced Explanation Evaluation Framework." *arXiv,* https://doi.org/10.48550/arXiv.2410.12855.

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. "Model Cards for Model Reporting." In *Proceedings of the Conference on Fairness, Accountability, and Transparency,* 220–29. https://doi.org/10.1145/3287560.3287596.

Mora-Reyes, Brigitte A., Jennifer A. Drewyor, and Abel A. Reyes-Angulo. 2025. "Advancing Equitable AI: Evaluating Cultural Expressiveness in LLMs for Latin American Contexts." *arXiv,* https://doi.org/10.48550/arXiv.2511.04090.

Mubarak, Hamdy, Abubakr Mohamed, and Majd Hawasly. 2025. "AraSafe: Benchmarking Safety in Arabic LLMs." In *Findings of the Association for Computational Linguistics: EMNLP 2025,* 9976–92. https://doi.org/10.18653/v1/2025.findings-emnlp.529.

Muhammad, Shamsuddeen Hassan, Idris Abdulmumin, Abinew Ali Ayele, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Saminu Mohammad Aliyu, Nelson Odhiambo Onyango, et al. 2025. "AfriHate: A Multilingual Collection of Hate Speech and Abusive Language Datasets for African Languages." *arXiv,* https://doi.org/10.48550/arXiv.2501.08284.

Ning, Zhiyuan, Tianle Gu, Jiaxin Song, Shixin Hong, Lingyu Li, Huacan Liu, Jie Li, et al. 2025. "LinguaSafe: A Comprehensive Multilingual Safety Benchmark for Large Language Models." *arXiv,* https://doi.org/10.48550/arXiv.2508.12733.

Ogundepo, Odunayo, Tajuddeen R. Gwadabe, Clara E. Rivera, Jonathan H. Clark, Sebastian Ruder, David Ifeoluwa Adelani, Bonaventure F. P. Dossou, et al. 2023. "AfriQA: Cross-Lingual Open-Retrieval Question Answering for African Languages." *arXiv,* https://doi.org/10.48550/arXiv.2305.06897.

Ojo, Jessica, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. "AfroBench: How Good Are Large Language Models on African Languages?" In *Findings of the Association for Computational Linguistics: ACL 2025,* 19048–95. https://doi.org/10.18653/v1/2025.findings-acl.976.

Ousidhoum, Nedjma, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. "Multilingual and Multi-Aspect Hate Speech Analysis." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),* 4675–84. https://doi.org/10.18653/v1/D19-1474.

Peppin, Aidan, Julia Kreutzer, Alice Schoenauer Sebag, Kelly Marchisio, Beyza Ermis, John Dang, Samuel Cahyawijaya, et al. 2025. "The Multilingual Divide and Its Impact on Global AI Safety." *arXiv,* https://doi.org/10.48550/arXiv.2505.21344.

Raji, Inioluwa Deborah, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. "AI and the Everything in the Whole Wide World Benchmark." In *Proceedings of the 35th Conference on Neural Information Processing Systems: Datasets and Benchmarks Track.* https://doi.org/10.48550/arXiv.2111.15366.

Ridoy, Shahriyar Zaman, Azmine Toushik Wasi, Koushik Ahamed Tonmoy, Taki Hasan Rafi, and Dong-Kyu Chae. 2026. "BengaliMoralBench: A Benchmark for Auditing Moral Reasoning in Large Language Models within Bengali Language and Culture." In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency,* 8051–88. https://doi.org/10.1145/3805689.3812230.

Roberts, Sarah T. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media.* Yale University Press. https://doi.org/10.12987/9780300245318.

Sawarkar, Piyush, Kundeshwar Vijayrao Pundalik, Ajay Nagpal, Viraj Thakur, Mohd Nauman, Shyam Pawar, Nihar Ranjan Sahoo, et al. 2026. "MILA (Multilingual Indic Language Archive): A Dataset for Equitable Multilingual LLMs." Preprint. Accessed August 7, 2026. https://openreview.net/forum?id=WPw6ERKUZL.

Shi, Enyi, Pengyang Shao, Yanxin Zhang, Chenhang Cui, Jiayi Lyu, Xiaobo Xia, Fei Shen, and Tat-Seng Chua. 2026. "Lingua-SafetyBench: A Benchmark for Safety Evaluation of Multilingual Vision-Language Models." *arXiv,* https://doi.org/10.48550/arXiv.2601.22737.

Verma, Sshubam, Mohammed Safi Ur Rahman Khan, Vishwajeet Kumar, Rudra Murthy, and Jaydeep Sen. 2025. "MILU: A Multi-Task Indic Language Understanding Benchmark." *arXiv,* https://doi.org/10.48550/arXiv.2411.02538.

Wynter, Adrian de, Ishaan Watts, Tua Wongsangaroonsri, Minghui Zhang, Noura Farra, Nektar Ege Altıntoprak, Lena Baur, et al. 2025. "RTP-LX: Can LLMs Evaluate Toxicity in Multilingual Scenarios?" *Proceedings of the AAAI Conference on Artificial Intelligence* 39 (27): 27940–50. https://doi.org/10.1609/aaai.v39i27.35011.

Ye, Meng, Karan Sikka, Katherine Atwell, Sabit Hassan, Ajay Divakaran, and Malihe Alikhani. 2023. "Multilingual Content Moderation: A Case Study on Reddit." In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics,* 3828–44. https://doi.org/10.18653/v1/2023.eacl-main.276.

Yong, Zheng Xin, Beyza Ermis, Marzieh Fadaee, Stephen Bach, and Julia Kreutzer. 2025. "The State of Multilingual LLM Safety Research: From Measuring the Language Gap to Mitigating It." In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing,* 15845–60. https://doi.org/10.18653/v1/2025.emnlp-main.800.

Yoo, Haneul, Yongjin Yang, and Hwaran Lee. 2025. "Code-Switching Red-Teaming: LLM Evaluation for Safety and Multilingual Understanding." *arXiv,* https://doi.org/10.48550/arXiv.2406.15481.

Yuan, Xiaohan, Jinfeng Li, Dongxia Wang, Yuefeng Chen, Xiaofeng Mao, Longtao Huang, Jialuo Chen, et al. 2025. "S-Eval: Towards Automated and Comprehensive Safety Evaluation for Large Language Models." *Proceedings of the ACM on Software Engineering* 2 (ISSTA): 2136–57. https://doi.org/10.1145/3728971.

Zhang, Wenjing, Xuejiao Lei, Zhaoxiang Liu, Meijuan An, Bikun Yang, KaiKai Zhao, Kai Wang, and Shiguo Lian. 2024. "CHiSafetyBench: A Chinese Hierarchical Safety Benchmark for Large Language Models." *arXiv,* https://doi.org/10.48550/arXiv.2406.10311.

Zhang, Zhexin, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2024. "SafetyBench: Evaluating the Safety of Large Language Models." In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 15537–53. https://doi.org/10.18653/v1/2024.acl-long.830.

Authors

Alisar Mustafa is the Head of AI Policy and Safety at Duco, where she partners with leading technology companies and governments to operationalize AI safety and governance. Her work focuses on translating complex policy commitments into robust technical implementations. Currently, her research explores the efficacy of AI safety benchmarks in low-resource languages, developing innovative methodologies to build stronger multilingual evaluations for high-risk safety topics that traditional, English-centric tools often overlook. With a background rooted in responsible innovation, Alisar previously co-founded and led AI policy and safety work at CivicSync and served on Meta's Responsible Innovation team. She also provided strategic counsel to the U.S. Census Bureau as a Fellow on the deployment of responsible AI. She authors the AI Policy Newsletter, providing critical analysis on emerging risks, global policy trends, and the future of AI safety. Alisar holds a Master of Science in Public Policy and Management from Carnegie Mellon University and a Bachelor of Arts in Political Science from San Francisco State University. Email: alisarmustafa2@gmail.com.

Cherry Wu is an independent AI policy and safety practitioner working at the intersection of evaluation and model governance. Her research focuses on multilingual AI safety, particularly performance gaps in low-resource languages, and her work on the multilingual gap in AI systems was presented at the 2025 Trust & Safety Professional Association APAC Summit and at a 2026 pre-conference event on AI and minority languages held ahead of the EU Presidency AI Summit. Her writing on AI policy has appeared through the Center for Security and Emerging Technology and the Centre for International Governance Innovation. She holds a Master of Science in Foreign Service from Georgetown University and a Bachelor of Arts in Political Science from McGill University. Email: cherry.sy.wu@gmail.com.

Positionality Statement

Alisar Mustafa is a native Arabic speaker and works in AI policy and safety, with direct experience building multilingual safety datasets in collaboration with language and domain experts, including prior applied work with Cherry Wu on human-authored data for safety fine-tuning. This experience, together with familiarity with Arabic and its dialectal variation, informed the review's attention to native authorship, to gaps between standard and colloquial varieties, and to the limits of translation-based evaluation.

Cherry Wu speaks Mandarin and a Chinese dialect, and works on AI evaluation and model governance. She has collaborated with East Asian dialect communities affected by data scarcity. Speaking both a high-resource standard variety and a dialect with far less representation in training data shaped the review's treatment of dialect and code-switching as core evaluation conditions rather than edge cases. Her dialect experience is nonetheless with a variety of a high-resource language, which differs in kind from the low-resource contexts the review centers.

The review is nonetheless conducted primarily from an English-language institutional and professional context, and the taxonomy reflects the evaluative priorities of Trust & Safety practice. For the many languages and communities in the corpus that the authors have no personal or professional connection to, the analysis rests on published benchmark documentation rather than direct community knowledge, which bounds what it can claim about how harm is experienced in those contexts.

Data Availability Statement

This article is a structured review of previously published safety benchmarks and did not generate new empirical data. All benchmarks analyzed are publicly available through the sources cited in the references, and the scoring rubric and complete per-benchmark scores are provided in Tables 1 and 2. The authors' scoring spreadsheet and inter-rater reliability data are available from the corresponding authors on reasonable request.

Funding Statement

The authors received no funding for this work.

Disclosure Statement

The authors declare no conflicts of interest.

Ethical Standards

This study is a structured review of previously published benchmarks and did not involve human subjects; institutional review board approval was therefore not required.

Keywords

Multilingual safety; large language models; benchmark evaluation; low-resource languages; content moderation.