Lost in Dialect: The Annotation Gap in Multilingual LLM Safety

Wajdi Zaghouani
Communication Program
Northwestern University in Qatar
Doha, Qatar
wajdi.zaghouani@northwestern.edu

Abstract

Large Language Models are increasingly used as safety infrastructure for detecting harmful online content and moderating social media across multiple languages. Yet their effectiveness remains uneven across linguistic communities. This disparity reflects not only disparities in training data availability but also structural problems in annotation design. We argue that a central source of multilingual safety failure lies in the annotation gap underlying existing hate speech datasets. Most annotation guidelines and safety benchmarks are developed for English and standard language varieties, overlooking dialectal variation and culturally embedded forms of hostility. Using Arabic dialectal discourse as a case study, we show how harmful speech expressed through dialects, sarcasm, code-switching, and culturally specific expressions often remains undetected by current annotation schemes. We introduce the concept of the Multilingual Safety Annotation Gap (MSAG), identifying four sources of bias: language coverage gaps, dialect representation gaps, cultural semantic gaps, and annotation guideline gaps. We discuss implications for LLM safety alignment and outline directions for culturally grounded multilingual annotation. This paper is primarily a conceptual and methodological position paper; rather than introducing a new benchmark or empirical evaluation, we aim to formalize the MSAG as a framework for analyzing systematic weaknesses in multilingual safety annotation pipelines.

1 Introduction

Large Language Models (LLMs) are rapidly becoming core infrastructure for moderating online discourse, detecting harmful language, and enforcing platform safety policies (Ouyang et al., 2022; Bai et al., 2022). Multilingual models such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and mT5 (Xue et al., 2021) demonstrate competitive performance across many natural language processing tasks, enabling cross-lingual transfer and multilingual deployment at unprecedented scale. Instruction-tuned and aligned variants of these models are now routinely deployed in content moderation pipelines spanning dozens of languages.

Despite these advances, substantial disparities persist in harmful speech detection across languages. Most datasets, benchmarks, and annotation frameworks were developed primarily for English (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018), producing safety models that perform unevenly across linguistic contexts. This unevenness reflects not only disparities in training data availability but also structural decisions made during dataset construction and annotation design. When labeling guidelines are developed primarily within English-language and Western cultural frameworks, they encode assumptions about harmful speech that may not generalize across languages and communities.

Arabic provides a revealing case study. While Modern Standard Arabic (MSA) dominates formal writing and broadcasting, most social media communication occurs in regional dialects, including Egyptian, Gulf, Levantine, and Maghrebi varieties, which differ substantially from both MSA and one another (Zaidan and Callison-Burch, 2014; Abdul-Mageed et al., 2021). These dialects include culturally embedded expressions, indirect insults, sarcasm, and code-switched content that frequently fall outside annotation frameworks designed primarily for standard language varieties. The result is a systematic blind spot in Arabic hate speech datasets and consequently in LLMs trained on those datasets.

We argue that multilingual safety failures stem not only from model limitations but also from weaknesses in the data annotation pipeline used to train safety systems. Annotation guidelines de-

Proceedings of the 1st Workshop on Multilinguality in the Era of Large Language Models (MeLLM 2026), pages 1–13

termine what counts as harmful speech, how it is labeled, and which linguistic expressions appear in training data. When these guidelines are built around English linguistic norms and standard language varieties, harmful discourse expressed through dialectal varieties can remain effectively invisible to automated systems and human reviewers unfamiliar with the cultural context.

We introduce the concept of the **Multilingual Safety Annotation Gap** (MSAG), defined as the mismatch between the linguistic diversity of harmful online discourse and the limited cultural and dialectal scope of existing annotation frameworks. Our contributions are threefold. First, we identify four structural sources of bias in multilingual hate speech annotation pipelines and for each propose concrete operational criteria to detect and measure its presence. Second, we illustrate these issues using examples from Arabic dialectal discourse, including dehumanizing animal metaphors, religious curses, sarcasm, dialectal insults, code-switched hostility, and indirect social insults. Third, we discuss implications for LLM safety alignment, particularly reinforcement learning from human feedback (RLHF), and propose directions for community-informed annotation frameworks. This paper is primarily a conceptual and methodological position paper; rather than introducing a new benchmark or empirical evaluation, we aim to formalize the MSAG as a framework for analyzing systematic weaknesses in multilingual safety annotation pipelines and to outline a concrete research agenda for addressing them. This work contributes to the multilingual LLM research agenda aimed at developing equitable language technologies for diverse global communities (Bommasani et al., 2021; Weidinger et al., 2021).

2 Related Work

2.1 Hate Speech Detection and Dataset Construction

Research on automated harmful speech detection has expanded considerably since the mid-2010s. Early influential work focused primarily on English-language social media. Davidson et al. (2017) introduced one of the first widely used hate speech datasets compiled from Twitter, establishing the important distinction between hate speech and merely offensive language. Waseem and Hovy (2016) developed an annotated dataset targeting racist and sexist content, demonstrating that surface-level indicators alone were insufficient for accurate detection. Founta et al. (2018) characterized the landscape of abusive language more broadly, showing that the phenomena encompass overlapping categories including toxicity, harassment, and hate speech, with important consequences for annotation design and inter-annotator agreement.

Subsequent work has foregrounded the difficulty of consistent annotation at scale. Vidgen and Derczynski (2019) argued that reliable abusive content detection requires careful conceptual modeling and internally consistent annotation definitions, a requirement that has proven difficult to satisfy when guidelines are applied across different cultural contexts. Zampieri et al. (2019) introduced a hierarchical taxonomy for offensive language detection that has been adopted in several multilingual shared tasks. Röttger et al. (2021) introduced HateCheck, a suite of functional tests designed to reveal model weaknesses that aggregate metrics obscure, identifying systematic failures in both academic and commercial hate speech detection systems. The benchmark effort HatEval (Basile et al., 2019) extended evaluation to Spanish and Italian, demonstrating that cross-lingual annotation and modeling remains substantially more challenging than monolingual approaches.

2.2 Annotation Bias and Annotator Subjectivity

A growing body of scholarship has documented the ways in which annotator identity shapes dataset labels for sensitive content. Sap et al. (2019) demonstrated that hate speech detection models trained on standard English datasets encode racial biases reflecting the demographic composition of annotation pools, systematically over-flagging content produced by African American speakers. Pavlick and Kwiatkowski (2019) provided theoretical grounding for understanding annotator disagreement, arguing that apparent label noise often reflects genuine linguistic ambiguity rather than annotation error. Gordon et al. (2022) offered empirical evidence that annotator cultural background and personal experience influence labeling outcomes on socially sensitive classification tasks.

These findings have motivated calls for more inclusive and diverse annotation practices. Aroyo and Welty (2015) proposed preserving annotator disagreement as an informative signal rather than collapsing it into majority-vote aggregation, an ap-

proach that can reveal systematic cultural differences in how harmful speech is perceived. Prabhakaran et al. (2021) demonstrated that annotator perspective can systematically shift classification labels for socially sensitive content, a problem that is particularly acute when annotators are recruited from populations that do not share the cultural background of the communities whose speech is being labeled.

2.3 Multilingual NLP, Cross-lingual Transfer, and Safety

Multilingual representation learning has made it possible in principle to extend safety systems beyond English at scale. Devlin et al. (2019) introduced mBERT, Conneau et al. (2020) improved upon this with XLM-R, and Xue et al. (2021) extended the paradigm to sequence-to-sequence models with mT5. Despite these advances, cross-lingual transfer for safety-critical tasks remains substantially more difficult than for standard NLP benchmarks. Montariol et al. (2022) showed that zero-shot cross-lingual transfer for hate speech detection is particularly challenging because the task involves both linguistic and cultural specificity, and proposed auxiliary multilingual task training as one mitigation strategy.

The problem of multilingual safety in LLMs has begun to receive dedicated attention. Weidinger et al. (2021) offered an influential taxonomy of risks from large language models that includes discrimination, exclusion, and toxicity as central concerns. Bommasani et al. (2021) documented the societal risks associated with foundation models more broadly. Work specifically examining safety across languages has found that LLMs may produce substantially higher rates of harmful outputs in non-English languages, reflecting both the English-centricity of safety training data and the underrepresentation of non-English languages in alignment procedures (Deng et al., 2023; Wang et al., 2023a). Wang et al. (2023b) constructed the XSafety benchmark covering 14 safety issue types across 10 languages and showed empirically that all major LLMs evaluated were significantly less safe in non-English contexts than in English, providing direct quantitative evidence of the disparity our paper addresses at the annotation level. Efforts to extend the reach of instruction-tuned models to underrepresented languages, such as the Aya Dataset initiative (Singh et al., 2024) spanning 65 languages, have similarly highlighted that safety and helpfulness gaps are closely linked to the linguistic composition of training and alignment data. These recent empirical results confirm that the annotation-level gaps identified in the MSAG framework have measurable downstream consequences for deployed systems.

2.4 Arabic Hate Speech and NLP

Arabic NLP presents distinctive challenges due to the diglossic relationship between MSA and spoken dialects (Ferguson, 1959) and the richness of Arabic morphology. Zaidan and Callison-Burch (2014) developed early resources for automatic Arabic dialect identification. Abdul-Mageed et al. (2021) demonstrated the effectiveness of dialect-aware pre-training for Arabic NLP with the ARBERT and MARBERT models. The broader challenge of dialectal NLP across multiple languages is surveyed by Joshi et al. (2025), who document systematic performance degradation for non-standard varieties across a wide range of tasks.

Work specifically targeting Arabic harmful speech has grown substantially in recent years. Mubarak et al. (2017) introduced one of the first Arabic abusive language datasets. Mulki et al. (2019) released L-HSAB, a Levantine Arabic Twitter dataset annotated for hate speech and abusive language. Abu Farha and Magdy (2020) presented the ArSarcasm dataset, demonstrating that sarcasm is extremely challenging to detect automatically, with the best models achieving an F1 score of only 0.46. Mubarak et al. (2021) produced a large-scale Arabic offensive language dataset with fine-grained tags. Mubarak et al. (2023) proposed a language-agnostic emoji-anchored collection method enabling broader dialectal coverage. Zaghouani et al. (2024a) constructed a multi-label Arabic hate speech dataset from 15,965 annotated tweets, and Zaghouani et al. (2024b) organized the FIGNEWS shared task on multilingual bias and propaganda annotation. Dialect-specific gaps are particularly acute for Maghrebi varieties: Boucherit and Abainia (2022) constructed a corpus for Algerian dialectal Arabic including Arabizi, and Lanasri et al. (2023) developed deep learning approaches for Algerian hate speech. Haj Ahmed et al. (2024) conducted a targeted sociolinguistic audit of Levantine Arabic datasets, empirically documenting dialectal skew and annotation guideline inadequacy. These efforts highlight significant remaining gaps in Arabic safety resources at the level of fine-grained dialectal and pragmatic

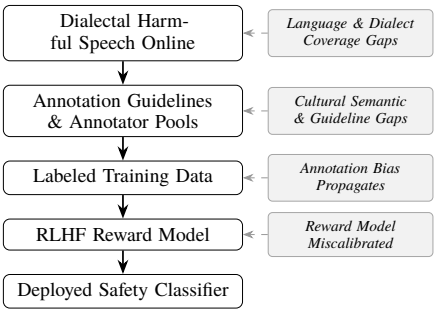

Figure 1: The MSAG compounds through the safety annotation pipeline. Each stage inherits and amplifies gaps introduced earlier, producing reward models and classifiers that provide unequal protection across dialect communities.

coverage.

3 The Multilingual Safety Annotation Gap

We define the **Multilingual Safety Annotation Gap** (MSAG) as the structural mismatch between the linguistic diversity of harmful online discourse and the limited cultural and dialectal scope of existing annotation frameworks. This gap reflects qualitative misalignments in how harm is conceptualized, operationalized, and measured across languages and cultures. Even if Arabic hate speech datasets contained as many examples as English datasets, they would still encode the MSAG if their annotation guidelines failed to capture the culturally specific forms that harmful Arabic speech takes.

The gap emerges from four interacting structural sources. These four dimensions are analytically distinct but operationally interdependent: a dataset may partially address one dimension while remaining deficient in others. For example, increased language coverage alone does not eliminate cultural semantic gaps if annotation guidelines remain grounded in Anglophone assumptions, and dialect representation gains are undermined when the guidelines used to label dialectal data were designed for standard varieties. Figure 1 illustrates how these gaps compound through the safety annotation pipeline.

3.1 Language Coverage Gap

The most visible dimension of the MSAG is the uneven distribution of annotated data across languages. Most hate speech and abusive language datasets have been developed for English (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018). Resources for other languages are considerably sparser, and for many languages with large online populations, purpose-built safety datasets do not exist. Multilingual safety models trained on English-dominant data rely on cross-lingual transfer to extend coverage, a process that degrades in proportion to the typological and script distance between source and target languages (Montariol et al., 2022).

Diagnostic criteria. A language coverage gap can be identified by (i) comparing the volume of purpose-built hate speech annotation across languages relative to speaker population size and social media activity, (ii) measuring cross-lingual transfer degradation for hate speech detection between English and the target language on held-out test sets, and (iii) auditing whether safety classifier training data includes native-language annotation or relies entirely on translated labels.

3.2 Dialect Representation Gap

Even within languages that have some annotated safety data, coverage is typically concentrated on standard written varieties. In the Arabic context, most available datasets draw primarily from MSA text, while the colloquial dialectal registers that dominate social media use remain underrepresented (Abdul-Mageed et al., 2021; Mubarak et al., 2021). This creates a gap that is qualitative as well as quantitative: dialectal speech encodes social meaning, identity, and relational dynamics in ways that differ systematically from formal registers. One promising collection strategy is the emoji-anchored approach of Mubarak et al. (2023), which exploits extralinguistic signals to surface harmful content across dialects without requiring predefined keyword lists. Such innovations can improve breadth of coverage but do not by themselves resolve the deeper problem of community-informed annotation guidelines.

Diagnostic criteria. A dialect representation gap can be identified by (i) computing the proportion of dataset tokens drawn from non-standard varieties versus MSA using automatic dialect identification tools, (ii) measuring performance degradation of a classifier trained on MSA-dominant data when evaluated on dialectal test sets, and (iii) auditing whether annotation guidelines explicitly cover dialectal lexical items, orthographic variants, and pragmatic forms.

3.3 Cultural Semantic Gap

Many forms of harmful discourse are culturally embedded in ways that resist straightforward cross-cultural translation. Animal metaphors used as insults carry different valences across cultures; in Arabic social media, such metaphors function as potent dehumanizing insults but may not be recognized as such by annotators working within a different cultural frame (Musolff, 2015). Religious expressions that function as curses in Arabic-speaking communities may not be recognized as harmful by annotators lacking the relevant cultural knowledge. Indirect speech acts targeting family honor require background cultural knowledge to decode their offensive intent (Haugh, 2015). Sarcasm and irony, which can convert seemingly positive statements into cutting attacks, require pragmatic inference that is difficult to operationalize in annotation guidelines oriented toward literal interpretation (Abu Farha and Magdy, 2020; Joshi et al., 2017).

Diagnostic criteria. A cultural semantic gap can be identified by (i) computing inter-annotator agreement separately for annotators from within and outside the target community, where systematic disagreement on specific expression types indicates culturally embedded phenomena, (ii) auditing whether annotation guidelines include examples of metaphor, sarcasm, and indirect speech in the target language, and (iii) testing whether classifiers trained on existing data fail on curated test sets of dialect-specific harmful expressions such as those in Table 1.

3.4 Annotation Guideline Gap

Annotation guidelines operationalize harm through explicit definitions, examples, and decision procedures. When these guidelines are developed primarily by researchers working within English-language and Western cultural frameworks, they encode implicit assumptions about the forms that harmful speech takes. Simplified binary or ternary classification schemes may fail to capture the gradients and contextual dependencies that characterize harmful expression in other languages. Guidelines that rely on keyword lists or surface-form indicators are particularly poorly suited to the indirect, metaphorical, and code-switched forms that harmful Arabic dialectal discourse frequently takes. The HateCheck framework (Röttger et al., 2021) revealed systematic weaknesses in state-of-the-art models precisely because standard held-out evaluation metrics obscure failures on functionally important cases, a problem compounded in dialect-aware and community-specific settings.

Diagnostic criteria. An annotation guideline gap can be identified by (i) auditing whether guidelines include culturally and dialectally specific examples alongside generic definitions, (ii) applying functional test suites such as HateCheck to measure failure rates on culturally specific categories not covered by existing guidelines, and (iii) measuring inter-annotator agreement on culturally ambiguous items as a proxy for guideline underspecification.

4 Harmful Dialectal Arabic: Illustrative Examples

Table 1 presents six categories of harmful Arabic dialectal expression that challenge conventional annotation frameworks. We constructed these examples to represent attested categories documented in the Arabic NLP literature (Mubarak et al., 2017; Mulki et al., 2019; Zaghouani et al., 2024a). We discuss each category in turn.

4.1 Animal Metaphors as Dehumanizing Insults

The expression انتو كلاب (“you are dogs”) and its dialectal variants exemplify how animal metaphors function as dehumanizing insults in Arabic. Animal metaphors have been documented as a widespread mechanism for dehumanizing outgroups across many languages (Musolff, 2015), but the specific animals involved and their severity vary considerably across cultural contexts. In Arabic social media, comparisons to dogs carry particular derogatory force and are frequently deployed to signal contempt across social, religious, or political boundaries.

4.2 Religious Curses

Expressions such as يوفقك لا الله (“may God not grant you success”) occupy an ambiguous position in Arabic harmful speech annotation. Invoking divine will to express ill intent is a recognized form of cursing in Arabic-speaking communities, understood as a hostile speech act that can escalate to stronger formulations wishing serious misfortune. Because religious language does not trigger the profanity and explicit slurs that many

Arabic Expression	Translation	Communicative Function	Dialect	Severity	Annot. Risk
انتو كلاب ما تفهمون شي	You are dogs, you understand nothing	Dehumanizing insult via animal metaphor	Gulf / Lev.	High	Underlabeled
الله لا يوفقك	May God not grant you success	Religious curse expressing ill will toward addressee	Pan-Arabic	Med.–High	Missed
ما شاء الله عبقري زمانك	God bless, the genius of your time	Sarcastic mockery of the addressee's intelligence	Pan-Arabic	Medium	Missed
يا غبي تفكر نفسك فاهم كل شي؟	You idiot, do you think you understand everything?	Direct dialectal insult with morphological variation	Egyptian / Gulf	Medium	Underlabeled
انتو ناس toxic والله	You people are toxic, I swear to God	Code-switched hostility (Arabic + English)	Lev. / Gulf	Medium	Inconsist.
واضح التربية ناقصة عندكم	It is clear that you lack proper upbringing	Indirect insult targeting family honor and social standing	Pan-Arabic	High	Missed

Table 1: Illustrative examples of harmful Arabic dialectal expressions and annotation challenges. *High* severity: serious personal attacks or dehumanization. *Medium*: harmful but potentially perceived as less severe by community-external annotators. *Underlabeled*: likely to receive less severe labels than warranted. *Missed*: likely to be labeled as non-harmful. *Inconsist.*: produces variable labels across annotators from different cultural backgrounds.

guidelines use as primary harm indicators, such expressions are frequently missed by automated systems and under-recognized by annotators unfamiliar with their pragmatic function (Mubarak et al., 2017).

4.3 Sarcasm and Ironic Praise

The expression ما شاء الله عبقري زمانك ("God bless, the genius of your time") illustrates the challenge that sarcasm poses for Arabic harmful speech annotation. Ostensibly positive formulations used ironically to mock a target are common in colloquial Arabic digital discourse and require pragmatic inference well beyond surface-level sentiment analysis. Abu Farha and Magdy (2020) showed that the ArSarcasm dataset, containing 10,547 tweets, proved extremely difficult for automated systems, with a best baseline F1 score of 0.46.

4.4 Direct Dialectal Insults

Direct insults in colloquial registers such as يا غبي ("you idiot") differ from their MSA counterparts in phonological form, orthographic realization, and contextual pragmatics. The same semantic content may appear in dozens of dialectal spelling variants in social media text, each with different detection profiles for keyword-based or embedding-based classifiers trained primarily on standard varieties. Abdul-Mageed et al. (2021) documented orthographic variability in dialectal Arabic as a major obstacle for NLP systems, and the same variability creates annotation inconsistency when annotators are not instructed to treat dialectal variants as equivalent.

4.5 Code-Switched Hostility

The expression انتو ناس toxic والله ("you people are toxic, I swear to God") combines Arabic dialectal syntax with an English loanword and an Arabic discourse particle. Code-switching between Arabic and English is extremely common in social media discourse produced by educated Arabic speakers in multilingual urban environments, particularly in Gulf, Levantine, and North African contexts. Annotation systems trained on monolingual data may fail to process code-switched text correctly, and the English element can distort cross-lingual transfer in unexpected ways.

4.6 Indirect Social Insults Targeting Family Honor

The expression واضح التربية ناقصة عندكم ("it is clear that you lack proper upbringing") exemplifies indirect speech acts that target family honor, a particularly sensitive dimension of social identity in Arabic-speaking societies. Remarks impugning someone's upbringing, family values, or social background function as serious insults within the relevant cultural context, carrying implications about family reputation that may not be evident

to annotators from outside that context (Haugh, 2015). Such expressions are systematically missed by annotation frameworks grounded in more explicit, keyword-oriented definitions of harmful speech.

5 Implications for LLM Alignment

5.1 RLHF and the Propagation of Annotation Errors

Recent LLM safety frameworks rely heavily on RLHF (Ouyang et al., 2022; Christiano et al., 2017). In standard RLHF pipelines, human annotators evaluate model outputs according to predefined safety guidelines, and these annotations are used to train reward models that shape model behavior during the alignment phase. The quality, cultural grounding, and linguistic coverage of the annotation process therefore directly determine the values encoded in the reward model.

The MSAG has a specific and consequential implication for RLHF: when annotators lack familiarity with dialectal language or community-specific harmful expressions, harmful content in those registers may be labeled as safe. Unlike random annotation noise, which can be partially mitigated by aggregation, culturally systematic misannotation is directionally biased. It consistently under-labels harm in certain communities' speech, producing reward models that fail to penalize specific forms of abusive language prevalent in non-English or dialectal contexts. To make this concrete: a reward model trained on annotations where an expression such as يوفقك لا الله is consistently labeled as neutral will not learn to assign a safety penalty to that expression regardless of how many annotators are consulted, because the error is not random but systematic. Such errors are detectable only through dialect-specific evaluation, precisely the function the Dialect-SafetyCheck framework introduced in Section 6 is designed to serve.

5.2 Anglophone Norm Encoding

The alignment procedures applied to many multilingual models may preferentially encode Anglophone norms of what constitutes harmful speech. Many RLHF datasets rely predominantly on English prompts and responses, and human annotators are predominantly recruited from English-speaking populations (Ouyang et al., 2022). Even when safety annotation is conducted in Arabic, it is frequently performed by annotators working in MSA or by bilingual annotators more comfortable with formal Arabic than with the full range of dialectal registers, leaving safety systems poorly calibrated for harm expressed through dialectal or community-specific means.

5.3 Downstream Disparate Impact

Communities whose languages and dialects are poorly represented in safety annotation may receive weaker protection from automated content moderation systems built on these aligned models. When a harmful insult in Gulf Arabic fails to trigger a safety classifier because the classifier was trained on data that systematically missed such insults, the speaker targeted by that insult is left without the protection that the safety system nominally provides. This represents a form of structural inequality embedded in the technical pipeline: the communities most likely to be targeted by harmful speech in underrepresented dialects receive the least effective automated protection, inverting the logic of safety infrastructure.

6 Toward Culturally Grounded Multilingual Safety Annotation

Addressing the MSAG requires coordinated intervention at multiple points in the annotation and alignment pipeline. We propose a research agenda organized around four priorities.

Dialect-aware annotation guidelines. Guidelines for Arabic hate speech annotation should be developed in close collaboration with sociolinguists and native speakers representing the major dialectal varieties, including Egyptian, Gulf, Levantine, and Maghrebi Arabic. Such guidelines should explicitly address the categories of harmful expression documented in this paper, including animal metaphors, religious curses, sarcastic praise, direct dialectal insults, code-switching, and indirect social insults. Severity calibration should be addressed explicitly: because perceived harm varies across dialect communities, as illustrated by the cross-dialect differences in Table 1, guidelines should include community-grounded severity rubrics rather than relying on universal thresholds. The adequacy of these guidelines should be validated empirically through inter-annotator agreement studies that include annotators from different regional backgrounds, with systematic analysis of disagreement patterns to identify sociolinguistically informed interpretation differences.

Representative annotator pools. Recruitment of annotators for Arabic safety datasets should prioritize regional and demographic diversity. Research by Gordon et al. (2022) and Prabhakaran et al. (2021) provides evidence that annotator background systematically influences labeling outcomes for sensitive content. For Arabic safety annotation, this means ensuring annotators represent the range of dialectal communities whose speech will be subject to safety systems, not only speakers of MSA or annotators recruited from one country or region. Following Aroyo and Welty (2015), annotator disagreement should be preserved as an informative signal rather than collapsed through majority voting, enabling downstream analyses of systematic cultural divergence in harm perception.

Dialect-SafetyCheck: A Blueprint for Functional Evaluation. The field currently lacks systematic benchmark datasets for evaluating hate speech detection across Arabic dialects at scale. We propose a *Dialect-SafetyCheck*, modeled on the HateCheck framework (Röttger et al., 2021), as a concrete operational target. Such a suite would consist of functional test cases organized by phenomenon (animal metaphors, religious curses, sarcasm, direct dialectal insults, Arabizi code-switching variants, and indirect honor-based insults), by dialect region (Egyptian, Gulf, Levantine, Maghrebi), and by severity tier, with each case carrying the Arabic expression, transliteration, dialect label, pragmatic category, and community-grounded severity rating. Evaluation should report performance separately for each dialect and pragmatic category, enabling researchers to track progress in closing the annotation gap per community.

Culturally informed annotation tools and documentation. Annotation interfaces and supporting documentation should provide annotators with cultural and pragmatic context relevant to the expressions they are labeling. This could include dialect glossaries explaining the pragmatic force of common expressions, guidance on indirect speech acts targeting honor and family reputation, and illustrative examples of sarcasm and irony as used in specific dialectal communities. Tools should also support Arabizi normalization so that Roman-script dialectal content is not filtered out at preprocessing. Such resources bridge the cultural knowledge gap for annotators who may be competent MSA speakers but less familiar with specific dialectal registers, and they also serve as documentation that makes annotation decisions more reproducible and interpretable across research groups.

7 Discussion

The MSAG has implications beyond Arabic. The four structural sources of bias, together with the diagnostic criteria and the pipeline model in Figure 1, are designed to be applicable wherever annotation guidelines and annotator pools do not reflect the full linguistic diversity of the communities they protect. Similar challenges arise in other linguistically diverse settings. Fayaz et al. (2025) introduced BIDWESH, a multi-dialectal Bangla hate speech dataset covering three regional varieties (Barishal, Noakhali, and Chittagong), motivated by the observation that existing resources fail to capture dialectal variation in harmful content. Their work illustrates a pattern that is general across low-resource and linguistically diverse languages: when annotation pipelines are designed without dialect-aware principles, harmful content may remain undetected. This pattern holds for Arabic, where dialectal variation is extensive but still underrepresented in annotation frameworks, and extends to a wide range of other language communities with complex dialect situations. Beyond dataset construction, the XSafety results of Wang et al. (2023b) provide direct empirical grounding for the MSAG at the model level: across four major LLMs tested on 10 languages, unsafe response rates were consistently and substantially higher in non-English settings. The fact that this disparity persists even in languages with moderate training data coverage such as Arabic strongly suggests that the problem is not reducible to data volume alone, consistent with our annotation-level analysis.

The MSAG framework also bears on ongoing efforts to improve multilingual alignment. Initiatives such as the Aya Dataset (Singh et al., 2024), which aggregated community-curated instruction data across 65 languages, have demonstrated both the feasibility and the difficulty of participatory multilingual data collection at scale. A parallel effort is needed specifically for safety annotation: community-sourced datasets that capture the pragmatic and cultural norms of dialect speakers, rather than projecting Anglophone-derived safety guidelines onto under-resourced linguistic contexts. The diagnostic criteria we propose for each MSAG dimension provide a structured starting point for au-

diting existing resources and prioritizing future community-informed annotation work.

The MSAG also has implications for how multilingual LLM safety is benchmarked. Current evaluations typically measure aggregate performance without disaggregating by dialect or register (Weidinger et al., 2021), obscuring the safety gap for specific communities. The Dialect-SafetyCheck proposed in Section 6 provides a mechanism for dialect-level disaggregation. We also recommend governance-level interventions: community juries from relevant dialect communities could calibrate severity ratings for culturally ambiguous expressions, and per-dialect safety cards published alongside model releases would make MSAG-relevant information accessible to practitioners. Documenting normative trade-offs between false positives that silence legitimate cultural expression and false negatives that permit genuine harm is essential for accountable deployment. Addressing the MSAG is ultimately a matter of linguistic equity: when safety systems fail to detect harmful discourse in underrepresented dialects, the communities most vulnerable to that harm bear the cost, inverting the very logic of safety infrastructure.

8 Conclusion

This paper introduced the Multilingual Safety Annotation Gap and argued that many multilingual LLM safety failures originate in dataset construction and annotation design rather than solely in model architecture or training data volume. Using Arabic dialectal discourse as a case study, we demonstrated how culturally embedded forms of harmful speech, including animal metaphors, religious curses, sarcasm, direct dialectal insults, code-switched hostility, and indirect social insults, can remain systematically invisible to annotation frameworks designed around English and standard language varieties. We identified four structural sources of the gap: the language coverage gap, the dialect representation gap, the cultural semantic gap, and the annotation guideline gap. For each, we proposed concrete diagnostic criteria to detect its presence in existing annotation frameworks, and we showed how these gaps compound through the alignment pipeline (Figure 1). We discussed how annotation-level gaps propagate through RLHF-based alignment to produce safety systems that provide unequal protection across linguistic communities. We also outlined a research agenda for developing more dialect-aware, community-informed multilingual safety annotation, including dialect-aware guidelines, representative annotator pools, a Dialect-SafetyCheck benchmark, and community-informed documentation tools. Closing this gap will require sustained collaboration between NLP researchers, sociolinguists, and the communities whose languages are at stake in the design of safety infrastructure.

Limitations

This paper is a position paper and does not present new empirical experiments or newly released datasets. The illustrative examples drawn from Arabic dialectal discourse were constructed to represent attested categories of harmful expression documented in the literature; they do not constitute a systematic corpus study. The categories discussed are not exhaustive: Arabic dialectal discourse contains additional phenomena including gendered insults, sectarian coded language, and politically inflected slurs that deserve separate treatment. We do not provide a quantitative operationalization of the MSAG. The severity ratings in Table 1 are based on the author's knowledge of the literature and community norms rather than on a systematic elicitation study. Our discussion of the implications for RLHF is necessarily general, given that the training data and annotation procedures used by major LLM developers are not fully disclosed. The proposed research agenda is programmatic rather than methodological, and decisions about guideline design, annotator recruitment, and benchmark construction will require empirical investigation beyond the scope of this paper.

Ethical Considerations

This paper discusses harmful speech directed at individuals and communities in Arabic-speaking contexts. The Arabic examples in Table 1 are constructed to represent categories of harmful expression for analytical purposes and do not correspond to data collected from specific real individuals. Research on harmful speech annotation inherently involves exposure of annotators to disturbing content. Any practical implementation of the recommendations in this paper should provide annotators with appropriate psychological support, fair compensation, and clearly articulated informed consent procedures. We recognize that decisions about what speech is counted as harmful involve political

and social choices with real consequences for affected communities, and we advocate strongly for community involvement in the design of annotation guidelines and for transparent documentation of the normative assumptions embedded in those guidelines. No new datasets were collected or released in connection with this paper.

Acknowledgment

This work was made possible by the National Priorities Research Program grant NPRPC14C-0916-210015 from the Qatar Development and Innovation Council (QRDI).

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Ibrahim Abu Farha and Walid Magdy. 2020. From Arabic Sentiment Analysis to Sarcasm Detection: The ArSarcasm Dataset. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools*, pages 32–39, Marseille, France. European Language Resource Association.

Lora Aroyo and Chris Welty. 2015. Truth is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Magazine*, 36(1):15–24.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:2212.08073*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63. Association for Computational Linguistics.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258*.

Oussama Boucherit and Kheireddine Abainia. 2022. Offensive Language Detection in Under-resourced Algerian Dialectal Arabic Language. *arXiv preprint arXiv:2203.10024*.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*, pages 512–515. AAAI Press.

Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and Yang Liu. 2023. MASTERKEY: Automated Jailbreaking of Large Language Model Chatbots. In *Proceedings of the 31st USENIX Security Symposium*. USENIX Association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Lin-*

guistics: Human Language Technologies (Volume 1: Long Papers), pages 4171–4186. Association for Computational Linguistics.

Azizul Hakim Fayaz, MD. Shorif Uddin, Rayhan Uddin Bhuiyan, Zakia Sultana, Md. Samiul Islam, Bidyarthi Paul, Tashreef Muhammad, and Shahriar Manzoor. 2025. BIDWESH: A Bangla Regional Based Hate Speech Detection Dataset. *arXiv preprint arXiv:2507.16183*.

Charles A. Ferguson. 1959. Diglossia. *Word*, 15(2):325–340.

Paula Fortuna and Sérgio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, 51(4):85:1–85:30.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media*, pages 491–500. AAAI Press.

Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19. ACM.

Michael Haugh. 2015. *Im/Politeness Implicatures*. De Gruyter Mouton, Berlin.

Ahmed Haj Ahmed, Rui-Jie Yew, Xerxes Minocher, and Suresh Venkatasubramanian. 2024. Navigating Dialectal Bias and Ethical Complexities in Levantine Arabic Hate Speech Detection. *arXiv preprint arXiv:2412.10991*.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J. Carman. 2017. Automatic Sarcasm Detection: A Survey. *ACM Computing Surveys*, 50(5):73:1–73:22.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2025. Natural Language Processing for Dialects of a Language: A Survey. *ACM Computing Surveys*, 57(6), Article 149. https://doi.org/10.1145/3712060.

Dihia Lanasri, Juan Olano, Sifal Klioui, Sin Liang Lee, and Lamia Sekkai. 2023. Hate Speech Detection in Algerian Dialect Using Deep Learning. *arXiv preprint arXiv:2309.11611*.

Syrielle Montariol, Arij Riabi, and Djamé Seddah. 2022. Multilingual Auxiliary Tasks Training: Bridging the Gap between Languages for Zero-Shot Transfer of Hate Speech Detection Models. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 347–363. Association for Computational Linguistics.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive Language Detection on Arabic Social Media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56, Vancouver, BC, Canada. Association for Computational Linguistics.

Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2021. Arabic Offensive Language on Twitter: Analysis and Experiments. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 126–135, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Hamdy Mubarak, Sabit Hassan, and Shammur Absar Chowdhury. 2023. Emojis as Anchors to Detect Arabic Offensive Language and Hate Speech. *Natural Language Engineering*, 29(6):1436–1457. Cambridge University Press.

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. L-HSAB: A Levantine Twitter Dataset for Hate Speech and Abusive Language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 111–118, Florence, Italy. Association for Computational Linguistics.

Andreas Musolff. 2015. Dehumanizing Metaphors in UK Immigrant Debates in Press and Online Media. *Journal of Language Aggression and Conflict*, 3(1):41–56.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training Language Models to Follow Instructions with Hu-

man Feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diab. 2021. On Releasing Annotator-Level Labels and Information in Datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138. Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional Tests for Hate Speech Detection Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection Using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10. Association for Computational Linguistics.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, et al. 2024. Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Bertie Vidgen and Leon Derczynski. 2019. Challenges and Frontiers in Abusive Content Detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. 2023. Do-Not-Answer: A Dataset for Evaluating Safeguards in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7137–7147. Association for Computational Linguistics.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael R. Lyu. 2023. All Languages Matter: On the Multilingual Safety of Large Language Models. *arXiv preprint arXiv:2310.00905*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93. Association for Computational Linguistics.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and Social Risks of Harm from Language Models. *arXiv preprint arXiv:2112.04359*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Salcianu, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498. Association for Computational Linguistics.

Wajdi Zaghouani, Hamdy Mubarak, and Md. Rafiul Biswas. 2024. So Hateful! Building a Multi-Label Hate Speech Annotated Arabic Dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15044–15055, Torino, Italia. ELRA and ICCL.

Wajdi Zaghouani, Mustafa Jarrar, Nizar Habash, Houda Bouamor, Imed Zitouni, Mona Diab, Samhaa El-Beltagy, and Muhammed AbuOdeh. 2024. The FIGNEWS Shared Task on News Media Narratives. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 530–547, Bangkok, Thailand. Association for Computational Linguistics.

Omar F. Zaidan and Chris Callison-Burch. 2014. Arabic Dialect Identification. *Computational Linguistics*, 40(1):171–202.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1415–1420. Association for Computational Linguistics.