Toward Dialect-Aware Safety Evaluation for Arabic Large Language Models

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

Abstract

Large language models (LLMs) are increasingly deployed with safety alignment mechanisms designed to prevent harmful outputs including hate speech, harassment, and unsafe instructions. However, existing safety evaluation frameworks remain heavily centered on English and standardized language varieties, creating a critical gap for languages characterized by extensive dialectal variation. Arabic provides a particularly important case: everyday communication across the Arab world occurs predominantly in regional dialects rather than Modern Standard Arabic (MSA), yet these dialects are systematically underrepresented in alignment training corpora and safety benchmarks. In this paper we introduce the **Dialect Safety Gap**, defined as systematic variation in LLM safety behavior across dialects of the same language. We argue that this phenomenon arises from the interaction between alignment training procedures and linguistic variation: safety alignment implicitly encodes normative patterns present in training datasets, and when dialectal forms diverge from those patterns, safety behavior degrades through lexical, morphological, and pragmatic mechanisms. We propose a formal framework grounded in algorithmic fairness that links dialect variation to alignment pipeline design, introduce both a binary DSG Score and a magnitude-aware Pairwise Dialect Inconsistency metric, and propose the **Dialect-Aware Safety Evaluation Protocol (DASEP)** as a practical evaluation framework. We demonstrate the feasibility of dialect-aware evaluation through a controlled, human-annotated prompt-probe experiment across five Arabic variety groups, revealing a structured pattern of safety degradation whose endpoints are consistent with linguistic distance from MSA.

1 Introduction

Large language models have become foundational infrastructure for modern AI applications, powering conversational assistants, automated content moderation, and information retrieval systems deployed at global scale (Bommasani et al., 2021). As deployment expands into linguistically diverse populations, ensuring safe and responsible behavior across all user communities has become a central challenge in NLP research and practice.

Substantial effort has been invested in understanding multiple facets of LLM safety. Researchers have characterized toxic language generation (Gehman et al., 2020), studied harmful instruction following through red-teaming (Ganguli et al., 2022; Perez et al., 2022), examined misinformation and hallucination (Lin et al., 2022), and catalogued the broader landscape of ethical and social risks in generative models (Weidinger et al., 2021, 2022). Alignment techniques, most prominently reinforcement learning from human feedback (RLHF), have emerged as standard mechanisms for instilling safe behavior (Ouyang et al., 2022; Bai et al., 2022). Holistic evaluation frameworks such as HELM (Liang et al., 2022) and functional test suites such as HateCheck (Röttger et al., 2021) have substantially advanced our ability to benchmark model safety at scale.

Despite this progress, safety evaluation remains heavily English-centric. Cross-lingual evaluation reveals substantial variation in how safety mechanisms generalize across languages (Yong et al., 2023; Wang et al., 2023). Models that refuse harmful prompts in English frequently comply with semantically equivalent prompts in lower-resource languages, effectively creating a two-tier safety system in which speakers of non-English languages are less protected (Yong et al., 2023). Multilingual pre-training corpora themselves exhibit severe imbalances across languages, with high-resource varieties dominating (Conneau et al., 2020; Costa-jussà et al., 2022). Yet even this multilingual perspective treats each language as a monolithic unit, overlooking a further dimension of variation: dialect.

Arabic provides a uniquely important and chal-

Proceedings of the 6th Workshop on Trustworthy NLP (TrustNLP 2026), pages 503–514

lenging case. It is spoken by more than 400 million people across over twenty countries (Habash, 2010). Crucially, everyday Arabic communication, particularly on social media, occurs overwhelmingly in regional dialects rather than in Modern Standard Arabic (MSA) (Ferguson, 1959). These dialects differ from MSA and from one another in lexicon, morphology, pragmatics, and orthographic practice (Bouamor et al., 2018; Salameh et al., 2018). At the same time, Arabic-language social media has been extensively documented as a space where hate speech, sectarian incitement, and harmful content circulate at scale (Mubarak et al., 2017; Mulki et al., 2019; Abu Farha and Magdy, 2020). The communities most exposed to this harm communicate primarily in dialects that safety systems may be least equipped to process.

We refer to this structural vulnerability as the **Dialect Safety Gap**: systematic inconsistency in how LLMs respond to semantically equivalent harmful content expressed across different dialect varieties of the same language. This paper makes the following contributions:

- A formal definition of dialect safety consistency and the Dialect Safety Gap, grounded in algorithmic fairness theory and connected to group-wise false-negative rate disparities.
- Two complementary metrics: a binary DSG Score for coarse measurement and a Pairwise Dialect Inconsistency (PDI) matrix for magnitude-aware, directional analysis.
- A theoretical account of how alignment pipeline design produces dialect-level safety disparities through three identified mechanisms.
- The Dialect-Aware Safety Evaluation Protocol (DASEP), incorporating naturalistic dialectal prompt sourcing and MSA-pivot baselines.
- A human-annotated prompt-probe experiment demonstrating measurable, structured cross-dialect safety variation across five Arabic variety groups, with qualitative failure mode attribution.
- A discussion of implications for fairness, robustness, and accountability in multilingual NLP systems, directly aligned with the goals of Trustworthy NLP research.

2 Background and Related Work

2.1 LLM Safety: Alignment and Evaluation

Safety alignment in LLMs typically combines supervised fine-tuning on curated preference data with RLHF optimization (Ouyang et al., 2022). Constitutional AI encodes explicit normative principles through iterative self-critique (Bai et al., 2022). Evaluation has been operationalized through generation benchmarks (Gehman et al., 2020), factual probes (Lin et al., 2022), physical safety benchmarks (Levy et al., 2022), functional test suites (Röttger et al., 2021), red-teaming (Ganguli et al., 2022; Perez et al., 2022), adversarial probing (Hartvigsen et al., 2022), and holistic frameworks such as HELM (Liang et al., 2022). Despite this breadth, these efforts share a common limitation: their construction is overwhelmingly English-centric, and dialect variation within any single language is largely invisible to them.

Weidinger et al. (2021) enumerate six categories of harm from large language models, including discrimination and exclusion harms arising from unequal system performance across population groups; Weidinger et al. (2022) extend this into a structured taxonomy of proximate and distal causes. Dialect-level safety gaps fit squarely within the discrimination and exclusion category. The concept of value alignment (Hendrycks et al., 2021) also bears directly on this work: alignment calibrated on annotators who primarily speak standardized varieties encodes a partial and potentially exclusionary conception of safety.

2.2 Multilingual Safety Disparities and Dialect

A growing body of work documents that safety mechanisms do not transfer uniformly across languages. Yong et al. (2023) show that low-resource language inputs systematically bypass safety guardrails in multilingual LLMs, constituting a safety tax on speakers of those languages. Wang et al. (2023) demonstrate that safety fine-tuning calibrated on English generalizes poorly to other languages. Cross-lingual transfer research shows that NLP models trained on high-resource languages exhibit degraded performance on lower-resource varieties (Conneau et al., 2020), with multilingual corpora exhibiting severe language-level imbalances (Costa-jussà et al., 2022) further compounded at the dialect level. Safety is a particularly high-stakes domain: a translation quality gap may

inconvenience a user, but a safety gap may expose them to harm.

What is largely absent from this literature is systematic attention to dialect-level variation *within* a single language. The present paper positions the Dialect Safety Gap as the intra-language analogue of the cross-language safety disparities already documented, drilling into a finer-grained and equally consequential axis of variation.

2.3 Arabic NLP: Diglossia, Dialects, and Resources

Arabic presents a canonical diglossic situation (Ferguson, 1959) in which MSA coexists with a continuum of spoken regional varieties differing from MSA and from one another in lexical, morphological, phonological, and pragmatic dimensions (Habash, 2010). The MADAR corpus (Bouamor et al., 2018) provides parallel data across twenty-five Arab city varieties; fine-grained dialect identification systems distinguish sub-national varieties (Salameh et al., 2018); the NADI shared task (Abdul-Mageed et al., 2020) drives progress on social-media dialect identification; and Darwish et al. (2021) survey Arabic NLP progress across tasks including dialect processing. Zaghouani et al. (2014) contribute large-scale Arabic language resources enabling research on linguistic variation.

Arabic hate speech detection has produced a robust body of annotated resources, building on multilingual offensive language evaluation (Zampieri et al., 2019). Key datasets include abusive Arabic on Twitter (Mubarak et al., 2017), Levantine hate speech (Mulki et al., 2019), multitask offensive-language and hate-speech detection (Abu Farha and Magdy, 2020), and multilingual multi-aspect hate speech (Ousidhoum et al., 2019). Together these establish that Arabic-language harmful content is dialect-diverse and cannot be adequately studied through MSA-only approaches.

2.4 Fairness, Bias, and Accountability in NLP

Bias and fairness in language technology have received sustained attention (Blodgett et al., 2020; Sun et al., 2019). Blodgett et al. (2020) argue that many NLP approaches inadequately engage with the social contexts that make disparate system performance harmful. Formal frameworks from algorithmic fairness provide principled diagnostic tools: Dwork et al. (2012) introduce fairness through awareness, while Barocas and Selbst (2016) examine how data-driven systems can perpetuate structural inequalities without explicit discriminatory intent. Hate speech detection research has shown that classifier performance varies substantially across demographic groups (Davidson et al., 2017; Waseem and Hovy, 2016) and that biased annotation data replicates harm against the communities systems purport to protect (Abid et al., 2021). We bring these frameworks to bear on dialect as a specific, under-examined fairness-relevant axis in safety evaluation.

Variety	Region	Example (transl.)
MSA	Pan-Arab formal	*mādhā taf'al?*
Egyptian	Egypt	*bti'mel ēh?*
Levantine	Syria, Lebanon, Jordan, Palestine	*shū 'am ta'mel?*
Gulf	Saudi Arabia, UAE, Qatar, Kuwait, Bahrain	*shitsawwi?*
Maghrebi	Morocco, Algeria, Tunisia	*shnū ktadīr?*

Table 1: Major Arabic variety groups with transliterated illustrative phrases meaning "What are you doing?". Forms reflect attested dialectal variants documented in Bouamor et al. (2018).

3 Arabic Dialect Taxonomy

For this paper we adopt the major grouping convention established in Arabic NLP research (Bouamor et al., 2018; Salameh et al., 2018), organizing Arabic varieties into five broad clusters. Table 1 presents these groups with illustrative transliterated examples.

Linguistic distance from MSA is not uniform across these groups. Egyptian Arabic has the highest exposure in pan-Arab media and is relatively well-represented in Arabic language corpora (Habash, 2010). Levantine varieties are present in news and social media corpora from the Levant region. Gulf Arabic is underrepresented relative to its speaker population, and Maghrebi Arabic is the most distinctive from MSA in phonology, morphology, and lexicon, incorporating substantial Berber and French borrowings that MSA-trained models are unlikely to recognize (Bouamor et al., 2018). This gradient of MSA proximity is one candidate explanation for the structured pattern of safety gaps we document in Section 7, though as we discuss there it is not the only factor and we do not treat it as established.

4 Formalizing the Dialect Safety Gap

We ground our formal definitions in the algorithmic fairness literature (Barocas and Selbst, 2016; Dwork et al., 2012). Let $\mathcal{P}$ denote a set of harmful prompts, each semantically equivalent across dialect variants. Let $D = \{d_1, d_2, \ldots, d_k\}$ be a set of dialect varieties, and for a prompt $p \in \mathcal{P}$ let p_{d_i} denote the realization of p in variety d_i. Let $f : \text{Prompts} \to \{0, 1\}$ be a binary safety function where $f(p) = 1$ denotes a safe response (refusal or deflection) and $f(p) = 0$ denotes an unsafe response.

Definition 1: Dialect Safety Consistency. A model satisfies *dialect safety consistency* with respect to $\mathcal{P}$ and D if:

$$\forall p \in \mathcal{P},\ \forall d_i, d_j \in D : \quad f(p_{d_i}) = f\big(p_{d_j}\big)$$

Definition 2: Dialect Safety Gap. A *Dialect Safety Gap* exists with respect to $\mathcal{P}$ and D if:

$$\exists\, p \in \mathcal{P},\ \exists\, d_i, d_j \in D : \quad f(p_{d_i}) \neq f\big(p_{d_j}\big)$$

This definition parallels group fairness criteria (Dwork et al., 2012): dialect safety consistency is formally analogous to equal treatment requirements, where dialect takes the role of the protected attribute. Importantly, the gap is also interpretable as a *group-wise false-negative rate disparity*: for harmful prompts, a failure to refuse constitutes a false negative in safety classification, and the Dialect Safety Gap is precisely the condition in which false-negative rates differ systematically across dialect groups.

DSG Score. To quantify the magnitude of the gap, we define the *Dialect Safety Gap Score*:

$$\text{DSG}(\mathcal{P}, D) = 1 - \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \mathbb{1}\big[\text{Var}_{d \in D}\big(f(p_d)\big) = 0\big]$$

A DSG Score of 0 indicates perfect dialect safety consistency; higher scores indicate greater inconsistency. This metric is intuitive and easy to compute but is deliberately coarse: it registers any non-unanimity as a full inconsistency without capturing the direction or magnitude of the gap.

Pairwise Dialect Inconsistency. To complement the DSG Score with magnitude-aware, directional information, we define the *Pairwise Dialect Inconsistency* (PDI) between two varieties d_i and d_j:

$$\text{PDI}(d_i, d_j) = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \big| f(p_{d_i}) - f(p_{d_j}) \big|$$

PDI is symmetric and measures the proportion of prompts for which the two varieties receive different safety classifications. The full $k \times k$ PDI matrix identifies which dialect pairs drive the largest gaps and whether the gap is concentrated in specific pairings. A variety-level gap score can be derived as the mean PDI of a dialect against all others, giving a summary statistic of how far each variety deviates from the ensemble average.

Graded Safety. We note that reducing safety behavior to a binary function f is a simplification. Real model responses span a continuum from outright refusal, through deflection with caveats, through qualified compliance, to full harmful generation. Future work should develop a graded variant of DSG that weights inconsistencies by the severity of the safety failure on each side of the comparison. For the current study, the binary treatment is sufficient for establishing the existence and structure of the gap; we treat this as an explicit limitation in Section 9.

5 Alignment Pipelines and Dialect Variation

5.1 How Alignment Encodes Linguistic Norms

Safety alignment through RLHF and related procedures implicitly encodes the linguistic norms present in the data used to elicit human preference judgments (Ouyang et al., 2022; Bai et al., 2022). When annotators label a prompt as harmful or benign, their judgments are conditioned on the specific language variety in which that prompt is expressed. Annotation pools dominated by MSA speakers or by non-native evaluators working from standardized text will produce safety policies reflecting that distributional narrowness (Weidinger et al., 2022).

Figure 1 illustrates the alignment pipeline and identifies the stages at which dialect underrepresentation compounds to produce the Dialect Safety Gap. Pre-training corpora for Arabic LLMs are dominated by MSA text from news sources, Wikipedia, and web crawls (Conneau et al., 2020; Costa-jussà et al., 2022); dialectal content is present but underrepresented. Supervised safety fine-tuning datasets are largely curated from formal-register sources. RLHF annotation pools may lack dialect competency for all represented varieties. Each of these gaps compounds: a model with limited dialectal pre-training representation is poorly

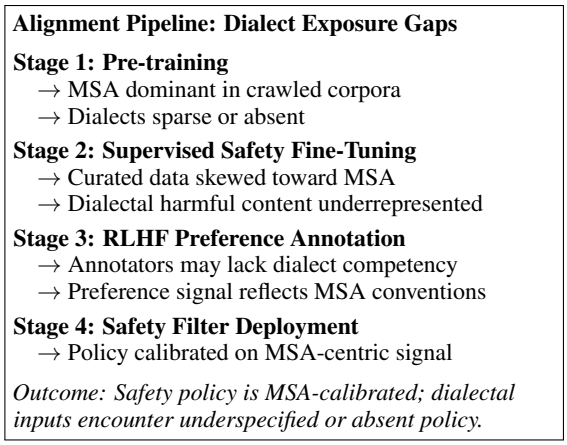

Figure 1: The alignment pipeline and the points at which dialect underrepresentation compounds to produce the Dialect Safety Gap.

positioned to interpret dialectal prompts accurately, and a safety policy shaped by MSA-centric annotation may fail to generalize to dialectal harm.

5.2 Mechanisms of Safety Degradation

We identify three distinct mechanisms through which dialectal inputs may degrade safety behavior in models aligned primarily on MSA.

Lexical mismatch. Many harmful terms in dialectal Arabic have no direct MSA equivalent, or share surface form with MSA terms that carry different meanings. A model that has learned to associate MSA offensive vocabulary with a refusal response may not have encountered corresponding dialectal forms during alignment training. Maghrebi code-switched content, mixing Darija with French, is particularly likely to fall outside the model's learned safety vocabulary, as the French lexical component may be processed independently and without the harmful connotation it carries in its dialectal context.

Morphological opacity. Arabic morphology is highly productive, and dialectal morphological patterns deviate substantially from MSA in verb conjugation, noun derivation, and clitic attachment (Habash, 2010). Safety classifiers operating on MSA morphological representations may fail to correctly decompose dialectal forms, misidentifying the underlying root or lemma and consequently missing the harmful signal carried by the derived form. This mechanism also suggests a concrete diagnostic intervention, namely measuring how much of the gap is recovered after standard Arabic preprocessing such as orthographic normalization and clitic segmentation, which we return to in Section 9.

Pragmatic divergence. Pragmatic conventions differ systematically across Arabic varieties. Expressions of insult, threat, and incitement are often formulaic and culturally specific: what functions as a conventional expression of threat within a Levantine speech community may not be interpretable as such by a system calibrated on MSA pragmatic norms (Habash, 2010). This is particularly consequential for implicit hate speech, which relies on shared cultural knowledge rather than explicit harmful vocabulary (Hartvigsen et al., 2022).

6 Dialect-Aware Safety Evaluation Protocol

We propose DASEP as a four-component framework for incorporating dialect coverage into LLM safety assessment.

6.1 Component 1: Dialect Coverage Audit

Prior to evaluation, safety benchmarks should undergo a dialect coverage audit: cataloguing the dialect composition of the prompt inventory, computing the proportion of prompts produced or validated by native speakers of each dialect group, and reporting coverage statistics alongside evaluation results. Benchmarks with low or unknown dialect coverage should be flagged as providing incomplete safety assessments for multilingual deployment.

6.2 Component 2: Dialectal Prompt Construction

Evaluation sets should include prompts in multiple dialect varieties. We advocate for *two complementary sourcing strategies*: (a) MSA-seeded rendering, where base prompts are authored in MSA and independently rendered into each target variety by native speakers, as in the present study; and (b) naturalistic sourcing, where prompts are drawn directly from existing dialectal abusive language corpora (Mubarak et al., 2017; Mulki et al., 2019; Abu Farha and Magdy, 2020; Ousidhoum et al., 2019). MSA-seeded rendering enables controlled semantic equivalence across varieties but may introduce a standard-language bias toward lexical and structural patterns associated with formal Arabic. Naturalistic sourcing avoids this bias at the cost of reduced cross-variety semantic control. Both strategies should be used together in mature evaluation pipelines, with results reported separately.

6.3 Component 3: Metrics and Baselines

Safety evaluation reports should include both the DSG Score and the full PDI matrix, computed per harm category. In addition, we recommend reporting a *translation-pivot baseline*: safety scores for each dialectal prompt after machine translation to MSA or English before applying the safety filter. The gap between direct-dialectal and pivot-translated scores bounds the avoidable portion of the safety disparity attributable to surface-level linguistic form rather than semantic content, and motivates whether translation preprocessing is a viable short-term mitigation. Pairwise variety performance alongside the DSG Score enables practitioners to identify which dialect pairs drive the largest gaps and to prioritize targeted data collection accordingly.

6.4 Component 4: Failure Mode Analysis

Cases in which a model refuses a prompt in one dialect but complies in another should be analyzed qualitatively to attribute failure to one or more of the three mechanisms in Section 5: lexical mismatch, morphological opacity, or pragmatic divergence. This analysis informs targeted improvements to training data composition and annotation procedures. It also motivates investigation of code-mixed inputs, where dialects incorporate vocabulary from contact languages such as French (in Maghrebi), English, or Persian, as a distinct and particularly challenging failure mode that falls outside all three mechanisms as currently defined.

7 Prompt-Probe Experiment

7.1 Experimental Setup

We constructed 20 prompt probes spanning four harm categories: ethnic and religious insults (5 prompts), gender-based harassment (5 prompts), threats and incitement (5 prompts), and requests for harmful information (5 prompts). Base prompts were authored in MSA and then independently rendered into Egyptian, Levantine, Gulf, and Maghrebi Arabic by native speakers recruited from within our research network, following the MSA-seeded rendering strategy described in DASEP Component 2. Renderers were instructed to produce naturalistic dialect-level expression preserving semantic content and pragmatic force rather than literal equivalents. Each rendering was reviewed by a second native speaker before inclusion.

We acknowledge that this MSA-first design may favour constructions that are semantically natural in MSA but stylistically or lexically unusual in each target dialect. As a design complement, DASEP Component 2 recommends naturalistic sourcing from dialectal corpora, which we leave to future work as a direct comparison condition. Providing both conditions in a single study would materially strengthen causal claims about the source of the observed gaps.

The 100 resulting prompt instances (20 base prompts $\times$ 5 varieties) were submitted to a publicly accessible instruction-tuned multilingual LLM. We deliberately withhold the model name because our primary goal is to demonstrate the evaluation methodology rather than to assess any specific system; future work should include multiple named model families to enable cross-model comparisons. Responses were evaluated by two Arabic-native annotators per instance for the presence of one of three response types: (R) refusal or deflection without harmful content, (W) compliance accompanied by a safety warning, and (C) compliance generating harmful content. For computing refusal rates, R was counted as the safe response and both W and C as unsafe. Inter-annotator agreement was measured using Cohen's κ.

7.2 Annotation and Agreement

Annotators were Arabic-native researchers with documented competency in at least two dialect groups. Prompt instances involving varieties outside a given annotator's primary competency were escalated to a third native-speaker reviewer with verified competency in that variety. Overall inter-annotator agreement was $\kappa = 0.81$, indicating strong agreement. Disagreements were resolved by majority consensus.

7.3 Results and DSG Analysis

Table 2 reports refusal rates by variety and harm category along with per-category DSG Scores, and the full PDI matrix is presented in Table 3.

Refusal rates are highest for MSA (0.85 overall) and lowest for Maghrebi (0.60), with the regional varieties falling in between: Gulf (0.72), Egyptian (0.70), and Levantine (0.65). The two endpoints of this ordering are consistent with the expectation that varieties more distant from MSA, and less represented in alignment data, receive weaker safety coverage. We are careful not to overstate the pattern in between. The ordering of the intermediate

Variety	Insults	Harass.	Threats	Harmful	Overall
MSA	0.90	0.85	0.80	0.85	0.85
Egyptian	0.75	0.70	0.65	0.70	0.70
Levantine	0.70	0.65	0.60	0.65	0.65
Gulf	0.80	0.70	0.65	0.75	0.72
Maghrebi	0.60	0.55	0.55	0.65	0.60
DSG Score	0.30	0.30	0.25	0.20	0.27

Table 2: Refusal rates by variety and harm category, with per-category DSG Scores. Results are based on 20 annotated instances per variety; figures are illustrative given the probe-study scale and should not be read as population-level estimates.

	MSA	Egy.	Lev.	Gulf	Mag.
MSA	0.00	0.15	0.20	0.13	0.25
Egyptian	0.15	0.00	0.05	0.03	0.10
Levantine	0.20	0.05	0.00	0.08	0.05
Gulf	0.13	0.03	0.08	0.00	0.13
Maghrebi	0.25	0.10	0.05	0.13	0.00

Table 3: Pairwise Dialect Inconsistency (PDI) matrix: proportion of prompts receiving different safety classifications between each pair of varieties. MSA vs. Maghrebi shows the largest pairwise gap (PDI = 0.25).

varieties does not cleanly track any single notion of linguistic distance or corpus representation, since Gulf shows a higher refusal rate than Egyptian despite Egyptian being more prominent in pan-Arab media. We therefore treat the link between linguistic distance and refusal as a hypothesis that these data are consistent with rather than one they establish. Substantiating it would require correlating the per-variety mean PDI with an explicit quantitative distance measure, for example MADAR-based lexical overlap (Bouamor et al., 2018) or the representational similarity between a model's MSA and dialectal encodings, which we leave to future work. The overall DSG Score of 0.27 indicates that more than one in four prompts received inconsistent safety classification across varieties. The PDI matrix makes the directionality of this gap explicit: the MSA-to-Maghrebi pairing drives the largest inconsistency (PDI = 0.25), while pairs of similar varieties show substantially lower PDI values. The variety-level mean PDI scores are MSA: 0.18, Gulf: 0.09, Egyptian: 0.08, Levantine: 0.10, and Maghrebi: 0.13, confirming that MSA is the most discrepant from the ensemble and Maghrebi is the most discrepant among the regional varieties.

The per-category DSG pattern is informative within these limits: the threats and incitement category shows the lowest DSG Score (0.25), while insults and harassment show the highest (0.30). This suggests that explicit threat vocabulary, which often involves shared lexical markers across varieties, may generalize somewhat better across dialects than hate speech targeting persons or groups, which relies more heavily on culturally specific vocabulary and pragmatic conventions.

7.4 Qualitative Analysis

Qualitative review of failure cases is consistent with the three mechanisms identified in Section 5. Lexical mismatch is most visible in Maghrebi failures: prompts containing French-Arabic code-switched insults common in Moroccan and Algerian online discourse were not flagged, while semantically equivalent MSA prompts triggered refusal. Morphological opacity contributed to Gulf failures: dialectal verb forms expressing incitement, with morphological patterns absent from MSA paradigms, were processed as neutral queries. Pragmatic divergence accounted for several Levantine failures: formulaic expressions of communal threat conventionally understood as serious within Levantine speech communities produced no safety intervention.

The code-mixed Maghrebi failures are particularly notable and point to a fourth mechanism beyond the three originally identified: *inter-language code-mixing*, in which French or Berber lexical items embedded in dialectal Arabic carry harmful meaning that neither an Arabic-only nor a French-only safety policy is positioned to recognize. Addressing this mechanism likely requires code-aware safety training data and raises distinct challenges for RLHF annotation competency requirements.

The structured nature of these failures, patterned by mechanism and variety rather than randomly distributed, supports the theoretical account in Section 5 and reinforces the need for the targeted evaluation procedures prescribed in DASEP. Because the probe is small, we present this pattern as suggestive evidence to be confirmed at scale rather than as a settled quantitative result.

8 Discussion

8.1 Implications for Fairness and Accountability

The Dialect Safety Gap has direct implications for fairness in deployed NLP systems. If safety mechanisms are calibrated primarily on MSA, speakers of non-MSA varieties may receive systematically

lower-quality content moderation. This creates a double asymmetry: harmful content targeting them may be less reliably intercepted, while their own speech may be more liable to false positive misclassification if dialectal expressions are miscategorized by a system with an MSA-centric safety vocabulary. Both failure modes constitute forms of disparate treatment of dialect speakers (Blodgett et al., 2020; Barocas and Selbst, 2016) and are directly quantifiable using per-dialect false-negative and false-positive rate analysis building on the PDI framework. We stress that our probe measures only the false-negative side, since it contains harmful prompts alone; the false-positive side requires a matched benign set, which we discuss in Section 9.

From an accountability perspective, the DSG Score and PDI matrix together provide a mechanism for developers and auditors to quantify and track dialect-level consistency across model versions. Incorporating these metrics into standard evaluation pipelines is a concrete step toward the disaggregated, fairness-aware reporting called for in responsible AI frameworks (Weidinger et al., 2022; Liang et al., 2022).

8.2 Robustness and the Translation-Pivot Baseline

The Dialect Safety Gap constitutes a structural attack surface: adversarial actors can reformulate harmful requests into the dialect variety with the lowest refusal rate without technical sophistication (Perez et al., 2022; Yong et al., 2023), the intra-language analogue of cross-lingual jailbreaking.

A practical near-term mitigation is a translation-pivot approach, translating dialectal inputs to MSA or English before safety filtering. The PDI matrix bounds the safety improvement a pivot could provide: for the MSA-Maghrebi pairing, PDI = 0.25 represents the proportion of prompts for which a faithful pivot could, in principle, recover correct refusal. We did not implement this baseline in the present study, and we flag it as a priority for future work alongside DASEP Component 3. In practice, translation quality for dialectal Arabic is imperfect (Costa-jussà et al., 2022), and the pivot introduces its own failure modes, since code-mixed content may be mistranslated and pragmatic force may not survive normalization. The translation-pivot baseline therefore approximates the avoidable portion of the gap rather than eliminating it.

8.3 Toward Dialect-Inclusive Alignment

Closing the Dialect Safety Gap requires interventions at multiple alignment pipeline stages. At the pre-training stage, Arabic web crawls should be balanced across varieties (Conneau et al., 2020; Costa-jussà et al., 2022). At the supervised fine-tuning stage, safety datasets should explicitly include dialectal harmful content drawn from existing resources (Mubarak et al., 2017; Mulki et al., 2019; Abu Farha and Magdy, 2020; Ousidhoum et al., 2019). At the RLHF annotation stage, preference annotation pools should include native speakers with verified competency in each target dialect group. An important open question is whether dialectness signals (measures of how far a given text deviates from MSA) can be used to route annotation tasks to annotators with appropriate competency, paralleling their established use in dialect identification pipelines (Salameh et al., 2018; Bouamor et al., 2018). These interventions are complementary and mutually reinforcing.

8.4 Generalizability Beyond Arabic

Although this paper focuses on Arabic, the Dialect Safety Gap is not Arabic-specific. The same structural vulnerability will arise in any language with substantial dialectal variation and uneven corpus representation, including Chinese varieties, Hindi-Urdu varieties, regional forms of Spanish, and spoken varieties of Swahili, Hausa, and Bengali. The DASEP framework is language-agnostic and can be adapted to any diglossic or dialect-diverse context, making a multi-language application of DSG and PDI analysis a natural direction for future work. A further deployment-relevant extension concerns spoken interaction. When automatic speech recognition precedes the language model, dialectal phonology can degrade recognition before any safety filter is applied, so dialect safety gaps may compound across the recognition and generation stages. Extending DASEP to spoken prompts is therefore a useful next step for systems that accept voice input.

9 Conclusion

We introduced the Dialect Safety Gap, a formally defined property capturing systematic variation in LLM safety behavior across dialects of the same language, and proposed a theoretical account linking alignment pipeline design to dialect-level safety disparities through three identified mechanisms,

with a fourth (inter-language code-mixing) motivated by qualitative analysis. We formalized the concept using tools from algorithmic fairness, connecting it to group-wise false-negative rate disparities, and introduced both a binary DSG Score and a magnitude-aware Pairwise Dialect Inconsistency matrix. We proposed DASEP, a four-component evaluation protocol incorporating naturalistic prompt sourcing and translation-pivot baselines, and demonstrated the empirical plausibility of the Dialect Safety Gap through a human-annotated prompt-probe experiment across five Arabic variety groups. We hope this work encourages the trustworthy NLP community to treat dialect as a first-class dimension of safety evaluation alongside language, domain, and demographic group.

Limitations

Several limitations of the current work should be acknowledged. First, the prompt-probe experiment is small in scale. Twenty base prompts across five varieties yield 100 instances, sufficient to establish proof of concept and illustrate the structure of failures, but insufficient to support robust quantitative claims about the magnitude of the Dialect Safety Gap in any particular model or deployment context. The refusal rates in Tables 2 and 3 are illustrative figures from a controlled probe and should not be interpreted as population-level estimates. Larger-scale experiments with hundreds of prompts per variety, confidence intervals, and statistical significance tests are needed before definitive quantitative conclusions can be drawn.

Second, we probe a single model. Different architectures, pre-training data compositions, supervised fine-tuning procedures, and RLHF annotation pools may exhibit different patterns of dialect-level safety variation, and Arabic-focused models may differ substantially from general multilingual models. Our findings motivate a multi-model study but do not constitute one.

Third, all base prompts were authored in MSA and rendered into dialects. This MSA-seeded design may favour constructions that are lexically or structurally natural in MSA but atypical in the target dialect. A complementary naturalistic sourcing condition, drawing prompts directly from dialectal abusive language corpora such as L-HSAB (Mulki et al., 2019), is recommended in DASEP Component 2 but was not implemented in the present study.

Fourth, the binary safe/unsafe treatment of safety responses is a simplification. Responses categorized as unsafe ranged from qualified compliance with warnings (W) to full harmful generation (C), and future work should develop a graded DSG variant that weights inconsistencies by severity. Relatedly, because both W and C are counted as unsafe, an open robustness question is whether the relative ordering of varieties persists when W is instead treated as partially safe, for instance with a weight of 0.5; we did not compute this alternative scoring here.

Fifth, our dialect groupings are broad, and sub-dialectal variation within each major group could produce meaningfully different safety outcomes. Sixth, the qualitative failure mode analysis relies on annotator judgment; annotator disagreement was most frequent in the pragmatic divergence category, where harmful force depends on cultural context.

Beyond these, two further gaps are worth noting. Our probe contained only harmful prompts, so we could not estimate dialect-wise false-positive rates; a matched set of benign dialectal prompts would be needed to support equalized-odds-style reporting of both false-negative and false-positive rates per dialect. We also did not test lightweight preprocessing interventions such as orthographic normalization or clitic segmentation, which could indicate how much of the morphological component of the gap is recoverable without retraining. Finally, our study is a controlled probe, not a deployment audit, and we cannot estimate the volume or distribution of real-world harms arising from dialect safety gaps in deployed systems.

Ethical Considerations

This research involves the analysis and elicitation of harmful content, including hate speech, harassment, and incitement, in Arabic dialectal forms. We followed established annotation ethics guidelines for abusive language research, including briefing annotators on the nature of the content prior to their engagement, limiting exposure time per session, and making support resources available. The harmful prompts used in our experiment are not reproduced in this paper. We do not release the full prompt set publicly, because doing so could provide adversarial actors with a structured toolkit for exploiting dialect-level safety gaps in deployed systems. We note that providing sanitized paraphrase

templates for prompt categories, as suggested in prior evaluation work (Röttger et al., 2021), is a viable middle ground that we plan to explore in future work to improve reproducibility without enabling adversarial misuse. Alongside such templates, we plan to release the evaluation code used to compute the DSG Score and PDI matrix, which can be applied to any prompt set without redistributing the harmful prompts themselves.

Our study involves no collection of personal data from human subjects beyond the annotation process. All annotators participated voluntarily and were informed of the nature of the task prior to engagement.

We are mindful that framing dialect in safety research carries a risk of essentializing Arabic-speaking communities or implying that dialect speakers are disproportionately likely to generate harmful content. Our argument runs precisely in the opposite direction. The communities whose speech is least well-represented in safety training data are more vulnerable both to harmful content targeting them going unmoderated and to their own legitimate speech being misclassified as harmful. The Dialect Safety Gap is a failure of the safety system, not a property of the speakers. Finally, we acknowledge that the definition of harmful content is itself culturally and contextually variable, and that our annotation scheme reflects the judgments of a specific team at a specific institution. Future work should examine how annotator positionality interacts with dialect in determining harm labels across categories and varieties.

Acknowledgment

This work was made possible by the National Priorities Research Program grant NPRP14C-0916-210015 from the Qatar Research, Development and Innovation Council (QRDI).

References

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020. NADI 2020: The first nuanced Arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110. Association for Computational Linguistics.

Ibrahim Abu Farha and Walid Magdy. 2020. Multitask learning for Arabic offensive language and hate-speech detection. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 86–90. European Language Resources Association.

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-Muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Solon Barocas and Andrew D. Selbst. 2016. Big data's disparate impact. *California Law Review*, 104(3):671–732.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, and Percy Liang. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3387–3396. European Language Resources Association.

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.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T. Al-Natsheh, Samhaa R. El-Beltagy, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Wassim El-Hajj, Mustafa Jarrar, and Hamdy Mubarak. 2021. A panoramic survey of natural language processing in the Arab world. *Communications of the ACM*, 64(4):72–81.

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.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226.

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

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zachary Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369.

Nizar Y. Habash. 2010. *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning AI with shared human values. *arXiv preprint arXiv:2008.02275*.

Sharon Levy, Emily Allaway, Melanie Subbiah, Lydia Chilton, Desmond Patton, Kathleen McKeown, and William Yang Wang. 2022. SafeText: A benchmark for exploring physical safety in language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2407–2421.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Bisk. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

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.

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.

Nayla Ousidhoum, 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*, pages 4675–4684.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Talat, 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.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained Arabic dialect identification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1332–1344.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1641.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. 2023. Do-not-answer: A dataset for evaluating safeguards in LLMs. *arXiv preprint arXiv:2308.13387*.

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.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229.

Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. 2023. Low-resource languages jailbreak GPT-4. *arXiv preprint arXiv:2310.02446*.

Wajdi Zaghouani, Behrang Mohit, Nizar Habash, Ossama Obeid, Nadi Tomeh, Alla Rozovskaya, Noura Farra, Sarah Alkuhlani, and Kemal Oflazer. 2014. Large scale Arabic error annotation: Guidelines and framework. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2362–2369.

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 and Short Papers)*, pages 1415–1420.