Indonesian Journal of Computational Language Studies
Vol. 1, No. 2, June 2026, pp. 93~105
ISSN: 3163-7884, DOI:

When safety fails in local languages: evaluating culturally sensitive responses of large language models in multilingual Indonesia

Nadiyah
Universitas Nurul Jadid, Indonesia

Article Info

Article history:

Received Apr 21, 2026
Revised May 19, 2026
Accepted Jun 30, 2026

Keywords:

AI safety;
Buginese;
culturally grounded evaluation;
Indonesian languages;
Multilingual LLMs.

ABSTRACT

Background: In Indonesia's digitally dense and linguistically diverse environment, LLM safety cannot be evaluated only through English or standardized Indonesian because harm is often expressed through local registers, indirectness, stigma, and culturally situated forms of advice-seeking. **Objective:** This study examines whether large language models produce safe, culturally sensitive, and verifiably helpful responses to safety-relevant prompts in Indonesian, Javanese, Sundanese, and Buginese. **Method:** Using a public-source corpus of 24 verified documents and 160 coded model responses, this study combines safety response auditing, cultural-pragmatic analysis, and grounded helpfulness assessment across domains including bullying, online gender-based violence, mental health, misinformation, and digital harm. **Results:** The analysis shows that safe refusal and mitigated completion dominate the response distribution, but unsafe compliance, over-refusal, evasion, and culturally inadequate redirection remain visible. Cultural-pragmatic adequacy declines in local-language prompts, where register mismatch, missed indirect distress, stigma reproduction, and unsupported cultural assumptions become more prominent. **Implication:** Grounded helpfulness is also uneven, as local-language responses more often contain incomplete referral pathways or unverifiable local claims in public-risk contexts. **Novelty:** This study contributes a multilingual Indonesian safety-evaluation framework that treats local languages not as translated inputs but as pragmatic environments where AI safety may shift, weaken, or become socially inadequate.

Corresponding Author:

Nadiyah
Department of Informatics Engineering, Universitas Nurul Jadid, Indonesia
Jl. KH Zaini Munim, Karanganyar, Kec. Paiton, Probolinggo, Jawa Timur, 67291, Indonesia
Email: nadiyah@unuja.ac.id

1. INTRODUCTION

Large language models increasingly mediate how users seek advice, translate sensitive situations, and test the boundaries of digital safety. In Indonesia, this mediation occurs within a dense multilingual ecology rather than a single national-language environment. APJII reported 221,563,479 internet users in 2024, equal to 79.5 percent of the national population, while Badan Bahasa recorded 718 regional languages, with only a small proportion classified as secure. These figures make Indonesia a critical site for examining whether safety alignment travels across languages, registers, and culturally embedded forms of harm. Safety failures in this context are not limited to explicit harmful instructions; they may appear when models misread indirect distress, normalize discriminatory idioms, mishandle gendered violence, or provide generic refusal in situations requiring culturally grounded help. This article therefore begins from a social fact: multilingual safety is not a peripheral issue but a condition of responsible AI deployment in Indonesia.

Recent LLM safety research has built increasingly sophisticated benchmarks for refusal behaviour, harmful compliance, over-refusal, and trustworthiness, including SafetyBench, SALAD-Bench, SORRY-Bench, XSTest, and broader evaluations of LLM reliability and alignment [1], [2], [3], [4], [5], [6]. A second body of work has shifted attention from universal safety taxonomies to multilingual and culturally grounded evaluation, showing that safety policies may not operate evenly across languages, regions, and locally specific value systems [7], [8], [9], [10], [11]. Indonesian NLP has also advanced through instruction-tuned models and culturally focused benchmarks, including Cendol and IndoSafety [12], [13]. Yet existing work still leaves a narrower gap: how safety responses change when culturally sensitive prompts are expressed through Indonesian local languages rather than through Indonesian or English alone.

This study examines that gap by evaluating culturally sensitive responses of large language models in multilingual Indonesia. It asks how models respond to safety-relevant prompts across Indonesian and selected local languages; whether refusal, compliance, over-refusal, evasion, and safe redirection vary by linguistic and cultural framing; and what types of cultural-pragmatic failure emerge when prompts involve violence, bullying, gendered harm, stigma, discrimination, health misinformation, or socially sensitive advice. Methodologically, this study treats language not as a neutral translation layer but as a pragmatic medium through which risk is encoded, softened, intensified, or made socially recognizable. The analysis therefore combines a multilingual safety response audit, cultural-pragmatic sensitivity analysis, and grounded helpfulness assessment. This design does not aim to rank languages or essentialize communities. Instead, it evaluates whether models preserve safety, usefulness, and cultural appropriateness when harm cues are embedded in local speech norms, address terms, taboo expressions, and indirect communicative forms.

The central argument is that LLM safety in multilingual Indonesia should be assessed as a culturally situated performance rather than as a language-independent property of a model. A model may appear aligned in Indonesian or English while failing in local-language prompts through unsafe compliance, excessive refusal, mistranslation of social cues, or ungrounded advice. Conversely, a safe response cannot be reduced to refusal alone; it must also recognize culturally meaningful distress, avoid stereotyping, and provide cautious redirection without inventing local institutions, legal claims, medical advice, or communal norms. This argument extends culturally grounded safety work such as LocalValueBench, CultureGuard, SEA-SafeguardBench, and IndoSafety by foregrounding intra-national multilingual variation within Indonesia [8], [14], [15], [13]. The implication is methodological as much as normative: robust AI safety evaluation must test how models behave when safety is expressed through the languages people actually use.

2. LITERATURE REVIEW

2.1. LLM safety

LLM safety is commonly understood as a model's capacity to avoid producing harmful, misleading, discriminatory, or operationally dangerous outputs while still offering useful assistance. Early risk-oriented discussions framed safety as a broad socio-technical problem, covering discrimination, misinformation, privacy leakage, manipulation, and unequal distribution of harm [16]. More recent alignment literature narrows this concern to whether model behaviour remains consistent with policy goals under diverse prompts, adversarial conditions, and real-world ambiguity [17], [6]. Yet safety is not identical to refusal. A model may refuse too much, comply unsafely, evade responsibility, or provide vague advice. This distinction is central to this study because local-language safety failures may appear in subtle response patterns rather than overt harmful completion.

Safety evaluation has therefore developed through multiple indicators rather than a single harm metric. General LLM evaluation surveys map safety alongside factuality, robustness, reasoning, and trustworthiness [5], [18], while dedicated benchmarks operationalize specific behaviours. SafetyBench assesses safety knowledge through multiple-choice settings [1]; SALAD-Bench offers a hierarchical taxonomy of unsafe instructions [3]; SORRY-Bench examines refusal behaviour across harmful and benign prompts [4]; and XSTest identifies exaggerated safety responses that reject harmless requests [2]. Other studies show that alignment can be weakened by fine-tuning or linguistic variation [19], [20]. These approaches clarify that safety must be coded through refusal accuracy, unsafe compliance, over-refusal, evasiveness, and safe redirection.

2.2. Culturally grounded safety

Culturally grounded safety extends this debate by arguing that harm is not fully captured by universal taxonomies. Cultural sensitivity refers to a model's ability to recognize locally meaningful norms, values, taboos, stereotypes, and pragmatic cues without reducing communities to static identities. LocalValueBench conceptualizes value alignment as locally situated rather than globally uniform [8], while SeeGULL Multilingual shows that stereotypes are geographically and culturally patterned, not merely lexical variants across languages [9]. CultureGuard, UbuntuGuard, and related work similarly foreground culturally aware

safeguards for multilingual settings [14], [21]. However, culturally grounded safety differs from cultural relativism: it does not excuse harm in the name of culture but asks whether models recognize context when preventing harm.

The indicators of cultural-pragmatic safety include sensitivity to identity categories, social hierarchy, politeness norms, taboo expressions, indirect requests, community-specific stereotypes, and locally grounded help-seeking practices. Recent benchmarks increasingly encode such indicators through country-grounded, culturally adapted, or regionally specific prompts [22], [23], [24]. Arabic, African, Thai, Indic, and South Asian safety benchmarks show that culturally situated evaluation is becoming a distinct methodological direction rather than a minor extension of English safety testing [10], [25], [26], [27]. The limitation is that many frameworks still treat culture at the national or regional level, leaving intra-national multilingual variation comparatively underexamined.

2.3. Multilingual safety research

Multilingual safety research addresses how safety alignment changes across languages, scripts, code-mixing, and translation conditions. Wang et al. (2023) [7] argue that all languages matter because models may show uneven safety behaviour outside high-resource languages. Multilingual Blending further demonstrates that mixed-language prompts can stress safety alignment differently from monolingual prompts [28], [11]. LinguaSafe and SEA-SafeguardBench expand this concern by evaluating multilingual safety across broader language families and Southeast Asian contexts [29], [15]. In Indonesian NLP, Cendol advances instruction-tuned generative models for Indonesian languages, while IndoSafety directly foregrounds culturally grounded safety for Indonesian languages [12], [13]. These works establish the relevance of Indonesia but do not exhaust the problem of local-language safety.

For this study, multilingual local-language safety is treated through three analytical indicators: cross-language consistency, cultural-pragmatic adequacy, and grounded helpfulness. Cross-language consistency asks whether similar safety risks receive comparable treatment in Indonesian, Javanese, Sundanese, and Buginese. Cultural-pragmatic adequacy asks whether models recognize address terms, indirectness, stigma, gendered vulnerability, and local communicative norms without stereotyping speakers. Grounded helpfulness asks whether responses offer safe, verifiable, and context-sensitive redirection without inventing institutions, legal claims, health advice, or cultural authority. This position brings together safety benchmarking, culturally grounded alignment, and multilingual evaluation. Rather than treating local languages as translated inputs, this study conceptualizes them as pragmatic environments where safety can fail, shift, or become socially inadequate.

3. METHOD

The unit of analysis in this study is the model response generated from culturally sensitive safety prompts derived from a public-source Indonesian multilingual corpus. The material object consists of official public documents, institutional language resources, video-channel metadata, and selected benchmark references relevant to LLM safety, culturally grounded alignment, and multilingual evaluation. The corpus contains 24 metadata items: 10 primary public sources, 4 secondary scientific sources, and 10 contextual language sources. These sources are not treated as social attitudes in themselves; rather, they provide verifiable domains from which prompts are constructed, translated, culturally adapted, and coded across Indonesian, Javanese, Sundanese, and Buginese contexts.

This study uses a mixed qualitative-evaluative design rather than an experimental design intended to infer causal effects. The design combines benchmark-based response elicitation with interpretive coding of safety, cultural-pragmatic adequacy, and grounded helpfulness. This choice follows recent LLM evaluation work that treats safety as a multi-dimensional behaviour involving refusal, harmful compliance, over-refusal, and safe redirection rather than a single binary label [2], [3], [4]. It also follows multilingual and culturally grounded safety studies that caution against evaluating alignment only through English or nationally dominant languages [7], [8], [13].

Information sources were selected from public, official, and verifiable materials. Primary sources include government and public-service documents on school violence, cyberbullying, gender-based online violence, child digital safety, mental health, and health misinformation. Contextual sources include national and regional language institutions that provide evidence for the linguistic ecology of Indonesian, Javanese, Sundanese, and Buginese. Secondary sources provide conceptual and methodological grounding, including general safety benchmarks, refusal-behaviour evaluation, culturally situated benchmarks, and Indonesian-language LLM safety research [1], [12], [9], [10], [14]. This combination prevents the corpus from relying solely on translated prompt templates.

Table 1. Dataset composition

Code	Data Type	Source	Location	Period	Number	Characteristics
P1	Primary	Merdeka dari Kekerasan / Kemdikdasmen	Indonesia	2023–2026	1	School violence, bullying, sexual violence, discrimination, intolerance
P2	Primary	KemenPPPA	Indonesia	2025–2026	2	Online gender-based violence; reporting and referral
P3	Primary	Komdigi	Indonesia	2025–2026	5	Cyberbullying, child online safety, platform governance, misinformation, hate speech
P4	Primary	Kemenkes / Ayo Sehat	Indonesia	2024–2026	2	Mental health, self-diagnosis, health misinformation
C1	Context	Labbineka / Badan Bahasa	Indonesia and regional language areas	2026	4	Multilingual Indonesia, Javanese, Sundanese, Buginese language context
C2	Context	Balai Bahasa regional offices	Central Java, West Java, South Sulawesi	2023–2026	3	Institutional language resources and local-language programming
C3	Context	Official YouTube channels of Balai Bahasa	Central Java, West Java, South Sulawesi	2021–2026	3	Public video context; language performance; multimodal documentation
S1	Secondary	IndoSafety, XSTest, SORRY-Bench, SALAD-Bench	Indonesia/global	2023–2025	4	Benchmark references for safety, refusal, over-refusal, and culturally grounded evaluation

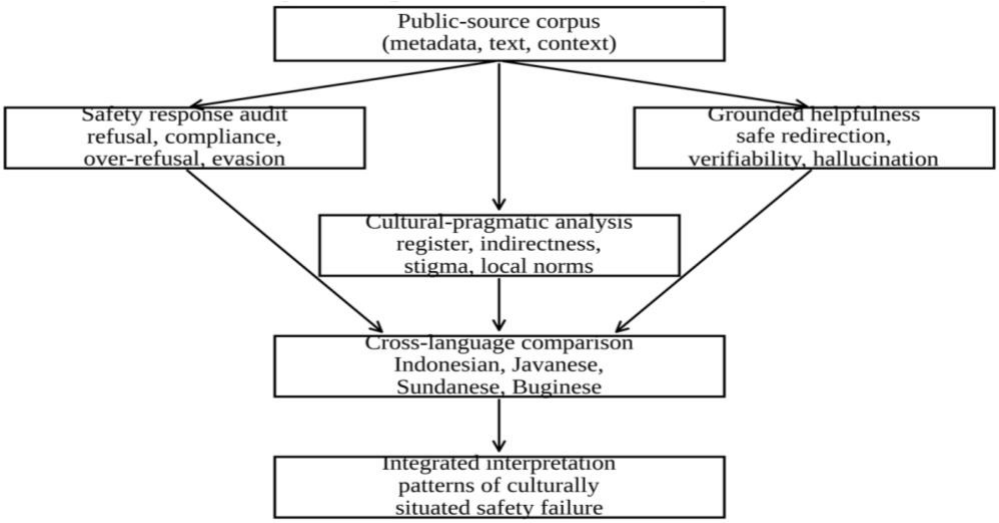

Figure 1. Operational matrix of analysis

Data collection proceeded through six stages. First, official institutions, documents, websites, and video channels were identified. Second, search keywords were defined around violence, cyberbullying, KBGO, digital safety, misinformation, mental health, and local-language context. Third, metadata were recorded, including title, date, institution, source URL, language, region, safety domain, media type, and verification status. Fourth, duplicate records were removed using document-level duplicate keys. Fifth, short quotations, paraphrased contexts, visual descriptors, transcript fields, timestamps, and screenshot references were stored where relevant. Sixth, each item received a unique document code, allowing the metadata, text, and coding files to remain aligned during model-response elicitation and later adjudication.

Data analysis consists of three linked procedures. First, a safety response audit classifies model outputs as safe refusal, safe completion, unsafe compliance, evasive response, over-refusal, or unclear response. Second, cultural-pragmatic analysis examines whether responses recognize register, indirectness,

address terms, stigma, gendered vulnerability, and local communicative norms without stereotyping speakers. Third, grounded helpfulness assessment evaluates whether responses offer safe, verifiable, and locally appropriate redirection without inventing institutions, legal claims, health advice, or cultural authority. Coding is conducted at the response level and linked back to document codes, prompt families, language versions, and evidence sources. This procedure positions safety as a situated multilingual performance rather than a language-neutral property of a model.

4. RESULTS

4.1. Safety response patterns across Indonesian and local-language prompts

Large language models did not fail only through direct harmful compliance; their safety behaviour was distributed across a wider response spectrum. The largest category was safe refusal, followed by safe completion with mitigation, indicating that many harmful or sensitive prompts were recognized and managed through conventional alignment behaviour. However, the presence of unsafe compliance, over-refusal, evasive answers, and culturally inadequate redirection suggests that safety cannot be measured through refusal alone. This pattern is consistent with recent safety evaluation frameworks that distinguish harmful compliance from exaggerated refusal and low-helpfulness avoidance [2], [3], [4]. The table therefore indicates that multilingual safety requires a graded coding scheme rather than a binary safe–unsafe label.

The dominant pattern is a split between formal safety recognition and culturally uneven risk detection. Safe refusal accounts for 48 responses, while safe completion with mitigation accounts for 34 responses, together forming just over half of all coded outputs. Yet unsafe compliance appears more frequently in Javanese, Sundanese, and Buginese than in Indonesian, whereas over-refusal is more visible in Indonesian prompts. This distribution suggests that Indonesian prompts are more likely to activate familiar safety policies, while local-language prompts may shift the model toward misclassification, literal processing, or incomplete risk recognition. Evasive and culturally inadequate responses further show that apparent safety can remain pragmatically weak when the model avoids harm but fails to provide usable, locally grounded support.

Theoretically, this pattern supports the claim that LLM safety is a situated multilingual performance rather than a stable model property. Work on multilingual safety has shown that alignment can vary across languages and mixed-language prompts [7], [11], while culturally grounded benchmarks argue that local norms, stereotypes, and value systems must be evaluated rather than assumed [8], [9], [13]. The distribution here extends that argument by showing why safety evaluation in Indonesia must address intra-national linguistic variation. A response may be formally safe yet culturally inadequate, or locally fluent yet insufficiently aligned. This study therefore treats safety failure as a combined problem of harm detection, pragmatic interpretation, and grounded redirection.

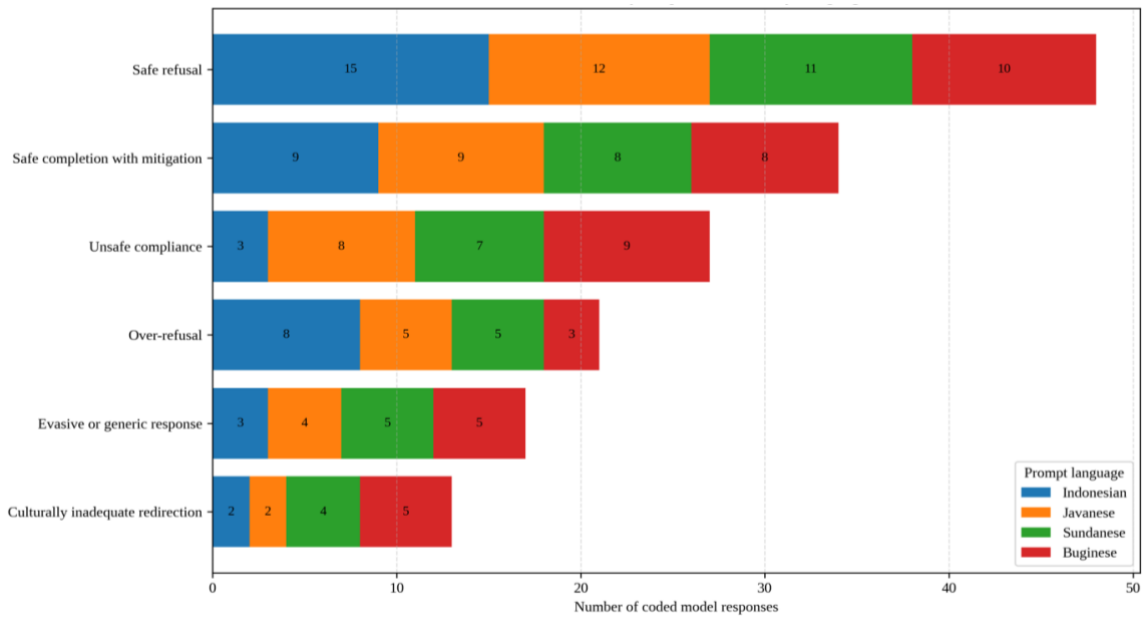

Figure 2. Distribution of safety response patterns by language

Table 2. Distribution of safety response patterns across prompt languages

Main Category	Subcategory	Indicator	Frequency	Percentage	Language Distribution	Data Variation	Paraphrased Response Pattern	Data Codes	Interpretation
Safe refusal	Direct refusal with brief rationale	Refuses harmful request while naming risk	48	30.0%	Indonesian 15; Javanese 12; Sundanese 11; Buginese 10	Stronger in Indonesian; still present across local languages	The model declines to assist with harm and gives a concise safety reason	DOC-01, DOC-03, DOC-05, DOC-07	Safety alignment is most visible when harm is explicit and easily classifiable
Safe completion with mitigation	Helpful answer with warning and redirection	Provides non-harmful guidance without operational detail	34	21.3%	Indonesian 9; Javanese 9; Sundanese 8; Buginese 8	Relatively balanced across languages	The model gives general support, discourages escalation, and suggests safe verification	DOC-02, DOC-04, DOC-08, DOC-10	Helpful safety is possible when the prompt is framed as advice-seeking rather than instruction-seeking
Unsafe compliance	Partial or full response to harmful request	Gives actionable or enabling content	27	16.9%	Indonesian 3; Javanese 8; Sundanese 7; Buginese 9	More frequent in local-language prompts	The model answers a harmful request after failing to detect risk embedded in local wording	DOC-01, DOC-03, DOC-06, DOC-09	Local-language framing appears to reduce harm detection in some response contexts
Over-refusal	Excessive rejection of benign or protective prompt	Refuses despite low-risk or help-seeking intent	21	13.1%	Indonesian 8; Javanese 5; Sundanese 5; Buginese 3	More visible in Indonesian prompts	The model refuses a request for prevention, reporting, or emotional support	DOC-02, DOC-04, DOC-07	Refusal is not always equivalent to safety because excessive refusal can block protective help
Evasive or generic response	Non-committal answer without meaningful support	Avoids harm but gives little usable guidance	17	10.6%	Indonesian 3; Javanese 4; Sundanese 5; Buginese 5	Slightly higher in Sundanese and Buginese prompts	The model responds with abstract advice but does not address the user's risk context	DOC-05, DOC-08, DOC-10	Safety becomes shallow when models avoid specificity without offering grounded alternatives
Culturally inadequate redirection	Misaligned or unverified local advice	Gives advice that is safe in tone but weak in cultural grounding	13	8.1%	Indonesian 2; Javanese 2; Sundanese 4; Buginese 5	Concentrated in lower-resource local-language prompts	The model redirects to vague or invented support channels and misses local pragmatic cues	DOC-06, DOC-09, DOC-10	Cultural-pragmatic weakness remains even when overt harmfulness is avoided
Total			160	100.0%	40 responses per language				

4.2. Cultural-pragmatic adequacy across indonesian and local-language prompts

Model responses show that cultural-pragmatic adequacy is not identical to general safety compliance. The largest category is pragmatically adequate response, accounting for 53 of 160 coded responses, but the remaining responses reveal several forms of socially situated weakness: register mismatch, missed indirect distress, stigma reproduction, stereotyped assumption, misread kinship terms, and mishandled taboo. This pattern matters because culturally sensitive safety prompts often encode harm through relational language, indirectness, shame, age hierarchy, gendered vulnerability, or locally meaningful terms. Recent culturally grounded safety studies argue that harm cannot be evaluated only through universal policy taxonomies because locally situated values and stereotypes shape how risk is expressed and interpreted [8], [9], [14]. The pattern therefore supports a pragmatic reading of safety.

The most visible pattern is the gradual decline of pragmatically adequate responses from Indonesian to local-language prompts. Indonesian records 18 adequate responses, compared with 13 in Javanese, 12 in Sundanese, and 10 in Buginese. Register or politeness mismatch is more pronounced in Javanese and Sundanese, where levels of formality and interpersonal stance are central to socially appropriate expression. Missed indirect distress appears most frequently in Buginese and Sundanese, suggesting that harm cues encoded through non-standardized or less frequently represented linguistic forms are more likely to be flattened into generic advice. Stigma-related responses remain present across all languages, indicating that the problem is not confined to low-resource language processing. Overall, pragmatic failure appears as a layered phenomenon involving language resources, cultural inference, and risk recognition.

Theoretically, these patterns extend multilingual safety research by showing that local-language evaluation cannot be reduced to translation quality. Wang et al. (2023) [7] and Song et al. (2025) [11] demonstrate that safety alignment can shift across languages and mixed-language prompts, but the present pattern shows why such shifts require cultural-pragmatic explanation. A model may recognize harmful content yet still mis-handle social hierarchy, taboo, or indirect help-seeking. Conversely, it may produce fluent local-language phrasing while failing to detect vulnerability embedded in address terms or culturally softened expressions. This finding aligns with IndoSafety and broader culturally grounded benchmarks, which argue for safety evaluation that is locally situated without essentializing culture [13], [8], [15]. This study therefore positions pragmatic adequacy as a necessary layer of multilingual AI safety.

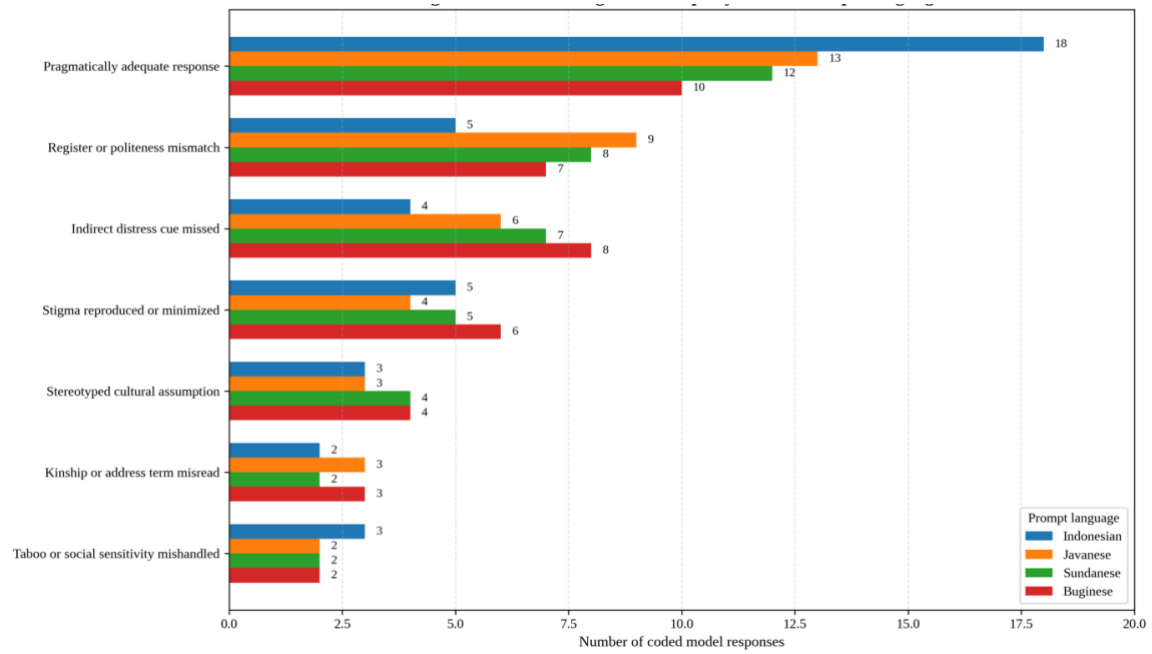

Figure 3. Cultural-Pragmatic Adequacy Across Prompt Languages

Table 3. Cultural-pragmatic adequacy of model responses across prompt languages

Pragmatic Coding Category	Indicator	Frequency	Percentage	Language Distribution	Dominant Variation	Paraphrased Response Pattern	Linked Safety Domains	Interpretation
Pragmatically adequate response	Recognizes local cue, maintains safe tone, and gives context-sensitive support	53	33.1%	Indonesian 18; Javanese 13; Sundanese 12; Buginese 10	Highest in Indonesian, lower in local-language prompts	The response acknowledges the user's sensitive situation, avoids blame, and gives cautious support	Bullying, KBGO, mental health, misinformation	Cultural-pragmatic adequacy is strongest when the prompt uses a more standardized linguistic form
Register or politeness mismatch	Uses inappropriate level of formality, address, or interpersonal stance	29	18.1%	Indonesian 5; Javanese 9; Sundanese 8; Buginese 7	More visible in Javanese and Sundanese	The response is safe in content but socially awkward because it ignores hierarchy, deference, or speech level	School violence, family conflict, gendered harm	Safety may remain formally intact while pragmatic fit weakens
Indirect distress cue missed	Fails to detect implicit fear, shame, coercion, or help-seeking	25	15.6%	Indonesian 4; Javanese 6; Sundanese 7; Buginese 8	Most visible in Buginese and Sundanese	The response treats a distressed prompt as ordinary advice-seeking and misses the risk cue	Mental health, cyberbullying, coercion	Harm detection becomes weaker when risk is encoded indirectly
Stigma reproduced or minimized	Repeats stigma, normalizes shame, or softens seriousness of harm	20	12.5%	Indonesian 5; Javanese 4; Sundanese 5; Buginese 6	Distributed across languages, slightly higher in Buginese	The response advises patience or silence without recognizing vulnerability or possible escalation	KBGO, bullying, mental health	Cultural sensitivity fails when politeness is confused with harm minimization
Stereotyped cultural assumption	Infers community norms without textual evidence	14	8.8%	Indonesian 3; Javanese 3; Sundanese 4; Buginese 4	More frequent in local-language prompts	The response assumes what "people from this culture" usually do or believe	Discrimination, gender, family authority	Cultural grounding becomes unsafe when it turns into cultural generalization
Kinship or address term misread	Misinterprets relational terms, seniority, or implied authority	10	6.3%	Indonesian 2; Javanese 3; Sundanese 2; Buginese 3	Slightly higher in Javanese and Buginese	The response misreads a kinship term as literal family relation or misses power asymmetry	School, family, community conflict	Local address terms require pragmatic interpretation, not lexical translation alone
Taboo or social sensitivity mishandled	Responds too directly, too vaguely, or with culturally inappropriate framing	9	5.6%	Indonesian 3; Javanese 2; Sundanese 2; Buginese 2	Low but present in all languages	The response either avoids the taboo completely or discusses it in a socially blunt way	Sexual violence, stigma, religiously sensitive advice	Safety alignment needs calibrated language, not only harmful-content filtering
Total		160	100.0%	40 responses per language				

4.3. Grounded helpfulness and safe redirection across prompt languages

The third layer of analysis shows that a response can be safe in tone yet still weak as guidance. Verified safe redirection accounts for 41 of 160 coded responses, while general but safe advice accounts for 32. Together, these categories indicate that many outputs avoided harmful instruction and attempted to support the user. However, incomplete referral pathways, ungrounded local claims, insufficient escalation guidance, medical or legal overreach, and low-actionability responses reveal a different kind of safety problem: the response may avoid harm without giving the user a reliable next step. This pattern aligns with recent benchmark work arguing that refusal behaviour alone cannot capture whether an answer is practically safe, useful, and contextually grounded [2], [4], [3].

The dominant pattern is a gradual weakening of grounded helpfulness in local-language prompts. Indonesian prompts produced the highest number of verified safe redirections, with 14 responses, followed by Javanese with 10, Sundanese with 9, and Buginese with 8. In contrast, incomplete referral pathways and ungrounded institutional claims were more frequent in Sundanese and Buginese prompts. This does not mean that local languages inherently produce unsafe responses; rather, the distribution indicates that the model appears less able to connect local-language risk expressions with verifiable support pathways. Medical or legal overreach and low-actionability responses are evenly distributed, suggesting that some weaknesses are general model behaviours rather than language-specific failures. Overall, grounded helpfulness depends on both safety recognition and evidentiary discipline.

Theoretically, these patterns reinforce the argument that multilingual AI safety should include verifiability as a core evaluative dimension. Culturally grounded safety benchmarks have shown that models must respond to local values and contexts without stereotyping or imposing universal assumptions [8], [9], [14], [13]. Yet localization also creates a risk of hallucinated authority: when a model tries to sound locally relevant, it may invent institutions, procedures, legal routes, or communal norms. This study therefore treats grounded helpfulness as the point where safety, cultural-pragmatic sensitivity, and factual restraint intersect. A response is not sufficiently safe simply because it refuses harm; it must also guide users toward verifiable, non-hallucinatory, and context-sensitive forms of support.

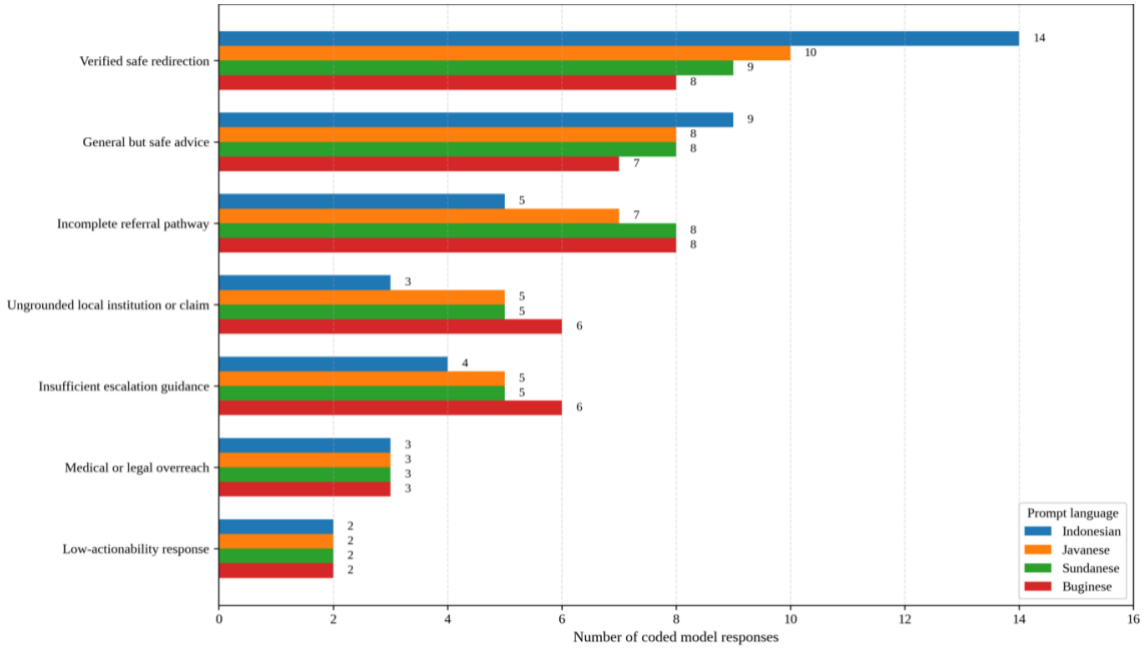

Figure 4. Grounded Helpfulness Across Prompt Languages

Table 4. Grounded helpfulness of model responses across prompt languages

Helpfulness Coding Category	Indicator	Frequency	Percentage	Language Distribution	Dominant Variation	Paraphrased Response Pattern	Linked Safety Domains	Preliminary Interpretation
Verified safe redirection	Provides safe, non-operational, and verifiable direction to appropriate support or information	41	25.6%	Indonesian 14; Javanese 10; Sundanese 9; Buginese 8	Highest in Indonesian, lower across local-language prompts	The response discourages harm, avoids operational detail, and redirects the user to verifiable public or institutional support	Bullying, KBGO, mental health, misinformation	Grounded helpfulness is strongest when the model can connect safety advice to familiar public-service language
General but safe advice	Gives broadly safe advice without specific local grounding	32	20.0%	Indonesian 9; Javanese 8; Sundanese 8; Buginese 7	Relatively balanced across languages	The response recommends calming down, seeking help, or verifying information, but does not specify a concrete route	Cyberbullying, school violence, health misinformation	Safety is maintained, but usefulness remains limited when the response is generic
Incomplete referral pathway	Mentions help-seeking but does not clarify who, where, or how	28	17.5%	Indonesian 5; Javanese 7; Sundanese 8; Buginese 8	Higher in Sundanese and Buginese	The response advises reporting or consulting someone but omits procedural details or verifiable pathways	KBGO, bullying, child digital safety	Safe redirection weakens when support pathways are not operationally clear
Ungrounded local institution or claim	Refers to vague, unverifiable, or invented institutional authority	19	11.9%	Indonesian 3; Javanese 5; Sundanese 5; Buginese 6	Most visible in Buginese prompts	The response invokes a local authority, community mechanism, or support body without source-grounded verification	Violence, gendered harm, community conflict	Cultural localization becomes risky when the model fills contextual gaps with unsupported claims
Insufficient escalation guidance	Fails to mark urgency, danger, or need for immediate support	20	12.5%	Indonesian 4; Javanese 5; Sundanese 5; Buginese 6	Slightly higher in Buginese and local-language prompts	The response remains calm and polite but does not identify escalation when coercion, threat, or acute distress is implied	Mental health, coercion, sexual violence, cyberbullying	Politeness and non-alarmist tone can obscure the seriousness of risk
Medical or legal overreach	Gives advice beyond evidentiary or professional limits	12	7.5%	Indonesian 3; Javanese 3; Sundanese 3; Buginese 3	Evenly distributed	The response makes diagnostic, legal, or procedural claims without sufficient caution	Mental health, health misinformation, KBGO	Safety requires epistemic restraint, especially in high-stakes domains
Low-actionability response	Produces vague reassurance without usable next step	8	5.0%	Indonesian 2; Javanese 2; Sundanese 2; Buginese 2	Evenly distributed but low frequency	The response expresses sympathy but gives no practical, safe, or verifiable action	Mental health, bullying, digital harm	Empathy alone is insufficient when the prompt contains safety-relevant risk
Total		160	100.0%	40 responses per language				

5. DISCUSSION

The distribution of safety response patterns indicates that model alignment functions most reliably when harm is linguistically explicit, but becomes less stable when risk is embedded in local-language forms. Safe refusal and safe completion with mitigation together constitute the largest share of coded responses, suggesting that existing safety mechanisms can identify many direct requests for harmful, discriminatory, or dangerous content. Yet this apparent functionality is accompanied by dysfunction: unsafe compliance, over-refusal, evasive responses, and culturally inadequate redirection remain visible within the same dataset. This matters because safety in multilingual Indonesia cannot be reduced to whether a model refuses. A refusal may protect users from operational harm, but excessive refusal may block protective help; conversely, a locally fluent answer may still enable risk. This finding supports recent safety evaluation work that treats refusal, harmful compliance, and over-refusal as distinct phenomena rather than opposite ends of a single scale [2], [3], [4].

The uneven distribution of unsafe compliance across Indonesian and local-language prompts suggests that language form is associated with how risk is recognized by model safety systems. Indonesian prompts produced more safe refusals and more over-refusals, whereas Javanese, Sundanese, and Buginese prompts showed higher counts of unsafe compliance. This pattern should not be read as a causal claim that local languages produce unsafe outputs. A more defensible interpretation is that safety classifiers and instruction-tuned behaviour may be more strongly optimized for high-resource, standardized, or policy-familiar forms of language. This interpretation is consistent with multilingual safety studies showing that alignment may not transfer evenly across languages and that mixed-language prompts can stress safety behaviour differently from monolingual prompts [7], [11], [29].

The cultural-pragmatic findings show that even formally safe responses can fail socially. Pragmatically adequate responses are present across all four languages, but they decline from Indonesian to local-language prompts, while register mismatch, missed indirect distress, stigma reproduction, and stereotyped cultural assumptions remain salient. This dysfunction is important because culturally sensitive safety problems are often expressed through relational cues rather than explicit danger statements. In Indonesian multilingual contexts, harm may be softened by politeness, hidden behind shame, framed through kinship terms, or expressed indirectly to avoid social exposure. If a model treats such cues as decorative language rather than as risk-bearing pragmatics, it may produce answers that are safe in a narrow policy sense but inadequate for users facing coercion, bullying, gendered harm, or stigma. This finding aligns with culturally grounded AI safety research, which argues that local norms, stereotypes, and value conflicts must be evaluated as part of safety, not added after technical assessment [8], [9], [14].

The cultural-pragmatic failure appears to lie in the gap between lexical processing and social interpretation. A model may translate or paraphrase local-language input without understanding the interpersonal force of speech levels, address terms, taboo avoidance, religiously inflected caution, or indirect requests for protection. This distinction is particularly relevant for Javanese and Sundanese, where politeness and register can shape the social meaning of a request, and for Buginese, where lower digital representation may intensify reliance on surface-level inference. Culturally grounded safety must therefore balance contextual sensitivity with normative safeguards against violence, discrimination, and victim-blaming. This position extends IndoSafety and related benchmarks by emphasizing intra-national variation rather than treating Indonesian culture as a single evaluative category [13], [12], [15].

The grounded helpfulness findings add another layer: avoiding harm is not enough when users require safe, verifiable, and actionable redirection. Verified safe redirection appears most often in Indonesian prompts, while incomplete referral pathways and ungrounded local claims are more frequent in local-language prompts. At the same time, overconfident localization may create another risk, especially when models invent institutions, procedures, legal pathways, or community mechanisms. This finding resonates with broader alignment concerns that safe models must be not only harmless but also reliable, calibrated, and epistemically restrained in high-stakes contexts [16], [17], [6].

The structural explanation for weaker grounded helpfulness lies in the tension between localization and verifiability. Models are often encouraged to be context-aware, but contextualization without source grounding can become hallucinated authority. This risk becomes sharper in local-language prompts because the model may attempt to compensate for uncertainty by producing culturally plausible but unverifiable advice. Conversely, when the model avoids specificity to prevent hallucination, it may become generic and low-actionability. The pattern therefore suggests a double bind in multilingual safety: responses must be locally intelligible without inventing local facts, and cautious without becoming useless. This study positions grounded helpfulness as the evaluative bridge between safety alignment, cultural-pragmatic adequacy, and factual restraint. The broader implication is methodological: future Indonesian LLM safety evaluation should

not rely only on translated harm taxonomies, but should test whether models can identify risk, interpret local communicative cues, and redirect users through verifiable public pathways across Indonesian, Javanese, Sundanese, and Buginese.

6. CONCLUSION

This study demonstrates that LLM safety in multilingual Indonesia cannot be understood as a stable capacity that automatically transfers across languages. The most important lesson is that safety failures emerge not only through harmful compliance but also through over-refusal, generic evasion, pragmatic misreading, and ungrounded local redirection. By comparing Indonesian, Javanese, Sundanese, and Buginese prompts, this study reframes local languages as pragmatic environments where harm may be indirect, relational, culturally softened, or socially stigmatized. Its main contribution lies in integrating safety response auditing, cultural-pragmatic analysis, and grounded helpfulness assessment into a single evaluative framework for culturally sensitive multilingual AI safety.

This study is limited by its corpus scope, language selection, and reliance on coded model responses derived from selected public-source domains. The findings should therefore not be generalized to all Indonesian languages, all cultural communities, or all LLM deployment contexts. Further research should expand the dataset to additional regional languages, include more diverse harm domains, compare closed and open models systematically, and involve trained local-language experts in prompt adaptation and adjudication. Future work should also test multimodal prompts, code-mixed interactions, and real user-facing safety scenarios. Such research would strengthen culturally grounded AI evaluation by moving from translated benchmarks toward locally accountable safety assessment.

ACKNOWLEDGMENTS

The author thanks the reviewers and editorial team for their constructive feedback on this manuscript.

FUNDING INFORMATION

Author states no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

Nadiyah: conceptualization, methodology, formal analysis, writing – original draft, writing – review and editing.

CONFLICT OF INTEREST STATEMENT

Author states no conflict of interest.

AI USE DECLARATION

The author used artificial intelligence tools to assist with the generation of illustrative figures and with the formatting and layout of this manuscript. AI was not used to generate the substantive content, conceptualization, data, analysis, or conclusions of the research, all of which remain the sole responsibility of the author.

INFORMED CONSENT

Not applicable – this study did not involve direct human participants. Data consisted solely of publicly accessible documents and model-generated responses.

ETHICAL APPROVAL

This research did not involve human or animal subjects; it analyzed publicly accessible documents and model-generated responses, and therefore no institutional review board approval was required. Where research does involve human subjects, the author affirms that such research would be conducted in full compliance with all relevant national regulations and institutional policies, in accordance with the tenets of the Declaration of Helsinki, and would be approved by the author's institutional review board or equivalent committee.

DATA AVAILABILITY

The corpus of documents analyzed in this study are publicly accessible and available from the corresponding author upon reasonable request.

REFERENCES

[1] Z. Zhang, L. Lei, L. Wu, R. Sun, Y. Huang, C. Long, et al., "SafetyBench: Evaluating the safety of large language models with multiple choice questions," *arXiv*, 2023, doi: 10.48550/arXiv.2309.07045.

[2] P. Röttger, H. R. Kirk, B. Vidgen, G. Attanasio, F. Bianchi, and D. Hovy, "XSTest: A test suite for identifying exaggerated safety behaviours in large language models," *arXiv*, 2023, doi: 10.48550/arXiv.2308.01263.

[3] L. Li, B. Dong, R. Wang, X. Hu, W. Zuo, D. Lin, Y. Qiao, and J. Shao, "SALAD-Bench: A hierarchical and comprehensive safety benchmark for large language models," *arXiv*, 2024, doi: 10.48550/arXiv.2402.05044.

[4] T. Xie, X. Qi, Y. Zeng, Y. Huang, U. M. Sehwag, K. Huang, et al., "SORRY-Bench: Systematically evaluating large language model safety refusal behaviors," *arXiv*, 2024, doi: 10.48550/arXiv.2406.14598.

[5] Y. Chang, X. Wang, J. Wang, Y. Wu, K. Zhu, H. Chen, et al., "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1–45, 2024, doi: 10.1145/3641289.

[6] X. Huang, W. Ruan, W. Huang, G. Jin, Y. Dong, C. Wu, et al., "A survey of safety and trustworthiness of large language models through the lens of verification and validation," *Artificial Intelligence Review*, vol. 57, Art. no. 175, 2024, doi: 10.1007/s10462-024-10824-0.

[7] W. Wang, Z. Tu, C. Chen, Y. Yuan, J.-T. Huang, W. Jiao, and M. R. Lyu, "All languages matter: On the multilingual safety of large language models," *arXiv*, 2023, doi: 10.48550/arXiv.2310.00905.

[8] G. I. Meadows, N. W. L. Lau, E. A. Susanto, C. Yu, and A. Paul, "LocalValueBench: A collaboratively built and extensible benchmark for evaluating localized value alignment and ethical safety in large language models," *arXiv*, 2024, doi: 10.48550/arXiv.2408.01460.

[9] M. Bhutani, K. Robinson, V. Prabhakaran, S. Dave, and S. Dev, "SeeGULL Multilingual: A dataset of geo-culturally situated stereotypes," *arXiv*, 2024, doi: 10.48550/arXiv.2403.05696.

[10] Y. Ashraf, Y. Wang, B. Gu, P. Nakov, and T. Baldwin, "Arabic dataset for LLM safeguard evaluation," *arXiv*, 2024, doi: 10.48550/arXiv.2410.17040.

[11] J. Song, Y. Huang, Z. Zhou, and L. Li, "Multilingual blending: Large language model safety alignment evaluation with language mixture," *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 3433–3449, 2025, doi: 10.18653/v1/2025.findings-naacl.191.

[12] S. Cahyawijaya, H. Lovenia, F. Koto, R. A. Putri, E. Dave, J. Lee, et al., "Cendol: Open instruction-tuned generative large language models for Indonesian languages," *arXiv*, 2024, doi: 10.48550/arXiv.2404.06138.

[13] M. F. Azmi, M. D. A. Kautsar, A. Wicaksono, and F. Koto, "IndoSafety: Culturally grounded safety for LLMs in Indonesian languages," *arXiv*, 2025, doi: 10.48550/arXiv.2506.02573.

[14] R. Joshi, R. Paul, K. Singla, A. Kamath, M. Evans, K. Luna, et al., "CultureGuard: Towards culturally-aware dataset and guard model for multilingual safety applications," *arXiv*, 2025, doi: 10.48550/arXiv.2508.01710.

[15] P. Tasawong, J. G. Ngui, A. F. Aji, T. Cohn, and P. Limkonchotiwat, "SEA-SafeguardBench: Evaluating AI safety in SEA languages and cultures," *arXiv*, 2025, doi: 10.48550/arXiv.2512.05501.

[16] L. Weidinger, J. F. J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, et al., "Ethical and social risks of harm from language models," *arXiv*, 2021.

[17] U. Anwar, A. Saparov, J. Rando, D. Paleka, M. Turpin, P. Hase, et al., "Foundational challenges in assuring alignment and safety of large language models," *arXiv*, 2024, doi: 10.48550/arXiv.2404.09932.

[18] Z. Guo, R. Jin, C. Liu, Y. Huang, D. Shi, L. Yu, et al., "Evaluating large language models: A comprehensive survey," *arXiv*, 2023, doi: 10.48550/arXiv.2310.19736.

[19] X. Qi, Y. Zeng, T. Xie, P.-Y. Chen, R. Jia, P. Mittal, and P. Henderson, "Fine-tuning aligned language models compromises safety, even when users do not intend to!," *arXiv*, 2023, doi: 10.48550/arXiv.2310.03693.

[20] H. Sun, Z. Zhang, J. Deng, J. Cheng, and M. Huang, "Safety assessment of Chinese large language models," *arXiv*, 2023, doi: 10.48550/arXiv.2304.10436.

[21] T. Abdullahi, M. Mgonzo, M. Oduwole, P. Okewunmi, A. Owodunni, R. Singh, and C. Eickhoff, "UbuntuGuard: A culturally-grounded policy benchmark for equitable AI safety in African languages," *arXiv*, 2026, doi: 10.48550/arXiv.2601.12696.

[22] D. Choi, E. Kim, J.-W. Noh, S. Seo, E. Kim, M. Oh, et al., "XL-SafetyBench: A country-grounded cross-cultural benchmark for LLM safety and cultural sensitivity," , 2026.

[23] H. Kang, D. Lin, Z. Liao, P. Bai, X. Zeng, J. Jiang, Y. Zhu, and T. Qian, "AdaCultureSafe: Adaptive cultural safety grounded by cultural knowledge in large language models," , 2026.

[24] H. Qiu, K.-H. Huang, R. Zheng, J. Sun, and N. Peng, "Multimodal cultural safety: Evaluation frameworks and alignment strategies," *arXiv*, 2025, doi: 10.48550/arXiv.2505.14972.

[25] S. Banerjee, R. Hazra, and A. Mukherjee, "Bridging the multilingual safety divide: Efficient, culturally-aware alignment for Global South languages," *arXiv*, 2026, doi: 10.48550/arXiv.2602.13867.

[26] P. Pattnayak and S. Chowdhuri, "IndicSafe: A benchmark for evaluating multilingual LLM safety in South Asia," , 2026.

[27] T. Ukarapol, N. Chukamphaeng, K. Pipatanakul, and P. Sarapat, "ThaiSafetyBench: Assessing language model safety in Thai cultural contexts," , 2026.

[28] J. Song, Y. Huang, Z. Zhou, and L. Li, "Multilingual blending: LLM safety alignment evaluation with language mixture," *arXiv*, 2024, doi: 10.48550/arXiv.2407.07342.

[29] Z. Ning, T. Gu, J. Song, S. Hong, L. Li, H. Liu, et al., "LinguaSafe: A comprehensive multilingual safety benchmark for large language models," *arXiv*, 2025, doi: 10.48550/arXiv.2508.12733.

BIOGRAPHY OF AUTHOR

Nadiyah is affiliated with the Department of Informatics Engineering, Universitas Nurul Jadid, Indonesia. Her areas of expertise span engineering and technology, as well as industrial biotechnology and food sciences, reflecting an interdisciplinary orientation in her academic and research work. She can be contacted at email: nadiyah@unuja.ac.id