

Multilingual Safety Is Model-Dependent: An Audit of Digital Mediation of Harm-Related AI Queries

Vivek Vaidya, Saubhagya Joshi, Vivek Singh

Rutgers University

vivek.vaidya@rutgers.edu, sau.joshi@rutgers.edu, v.singh@rutgers.edu

Abstract

Large language models are increasingly integrated into online information environments, including contexts where people seek information about sensitive and potentially harmful topics related to mental health. As these systems mediate access to such content, their safety mechanisms become part of the digital conditions shaping user risk and protection. In this paper, we use a diagnostic audit to examine how multilingual safety operates within this mediated environment. Using a two-stage, poem-based obfuscation method, we analyze responses from three frontier language models to harm-related queries in English, Spanish, and Hindi across three domains, including self-harm. The audit suggests that safety outcomes do not follow a simple hierarchy in which lower-resourced languages are consistently less protected. Instead, relative levels of protection vary by both language and model, with patterns that shift across systems and harm categories. This indicates that multilingual safety reflects model-specific alignment and design choices rather than inherent linguistic resource differences. From a digital mental health perspective, this finding has important implications for equity. Users seeking information during vulnerable moments may encounter different levels of protection depending not only on the language they use, but also on which AI system mediates their access. We argue that understanding mental health risk in online environments requires examining LLM safeguards as part of a broader ecosystem of digital mediation, moderation, and risk governance.

Introduction

Large language models are increasingly embedded in digital systems that mediate how people access information online. Beyond standalone chatbots, LLMs are integrated into search engines, browsers, and other interfaces that shape what information is surfaced, contextualized, or withheld, including for sensitive or potentially harmful topics. This shift is consequential for digital mental health research, as individuals often turn to online systems during vulnerable moments to seek information about self-harm, violence, or substance use.

The Digital Minds workshop emphasizes that mental health impacts arise from how sociotechnical systems structure access, interpretation, and risk. Prior work shows that

online platforms influence mental health through moderation practices, recommendation systems, and community norms, shaping exposure to harm and access to support. Research in social computing documents how self-harm related content circulates online, how users seek and provide support, and how moderation interventions affect well-being (De Choudhury et al. 2016; Chancellor et al. 2019; Jhaver, Bruckman, and Gilbert 2019).

As LLMs increasingly function as intermediaries between users and online information, their safety mechanisms become part of this broader ecosystem of digital mediation. Unlike traditional platforms, LLMs dynamically generate responses, making safeguards a central site where decisions about harm, protection, and access are enacted. From a mediation perspective, these safeguards operate as governance mechanisms that shape how risk is managed across contexts and populations.

Language is a critical but understudied dimension of this process. Multilingual LLMs are often framed as offering uniform access, yet their training data, alignment procedures, and safety tuning vary substantially. Social computing research shows that moderation systems can produce uneven effects across communities (Jhaver, Bruckman, and Gilbert 2019; Matias 2019). If LLM safeguards behave differently across languages, then users may experience different levels of protection depending on both language and system.

In this paper, we examine multilingual safety as a question of digital mediation rather than model performance. We conduct a diagnostic audit of three frontier language models responding to harm-related queries in English, Spanish, and Hindi across violence, drugs, and self-harm. Using a two-stage, poem-based obfuscation method, we probe how safeguards operate under indirect and creatively phrased requests.

Our findings show that multilingual safety is model-dependent rather than following a fixed hierarchy by language. Relative levels of protection shift across systems and harm categories, indicating that safety reflects model-specific alignment and design choices. From a digital mediation and mental health equity perspective, this variation matters because LLMs increasingly act as gatekeepers to harm-related information.

Contributions. This paper makes three contributions. First, we introduce a diagnostic audit framework for examining multilingual safety in LLMs as a form of digital mediation of harm-related information. Second, we show that safety outcomes vary by both language and model. Third, we situate multilingual LLM safety within broader discussions of digital mediation and health equity, arguing that AI safeguards should be studied alongside other mechanisms of moderation and risk governance.

Background and Related Work

This work builds on research examining how AI systems mediate access to harmful information and how safety mechanisms behave across languages and prompting strategies. Prior studies show that LLM safeguards often vary across languages, shaping how harm-related content is filtered or disclosed to users. Multilingual jailbreak evaluations report increased vulnerability outside English (Yong, Menghini, and Bach 2024; Deng et al. 2024), with code-switching further weakening protections across systems (Yoo, Yang, and Lee 2025). Audits and surveys highlight an English-centric focus in safety datasets and evaluation practices, raising concerns about uneven protection in multilingual settings (Yong et al. 2025). Related robustness analyses find variation across languages and across harm categories (Ji et al. 2023; Joshi et al. 2025), but typically assume that vulnerability follows a stable hierarchy tied to language resources.

A complementary line of work studies how obfuscation alters the digital mediation of harm by changing the surface form of requests. Prior work documents the effectiveness of camouflaged prompts, multi-turn strategies, and multimodal attacks in bypassing safeguards (Zheng, Zandsalimy, and Sushmita 2025; Mustafa et al. 2025). More recent studies show that stylistic transformations, including rewriting harmful requests as poems, substantially increase attack success, exposing limits in how alignment systems detect and moderate risk (Bisconti et al. 2025). Large-scale evaluations further suggest that such jailbreaks remain accessible despite ongoing safety improvements (Shen et al. 2024; Pathade 2025).

We combine multilingual analysis with poem-based obfuscation to audit how different models mediate harm-related queries across languages. This approach reveals model-dependent shifts in relative protection, reframing multilingual safety as a property of alignment and governance choices rather than language alone. In doing so, we situate LLM safeguards as part of the broader digital mediation infrastructure through which harm and protection are unevenly distributed.

Methods

Dataset, Languages, and LLMs

We evaluate three frontier LLMs: GPT-5, DeepSeek 3.2, and Gemini 2.5 using the dataset from (Joshi et al. 2025), which consists of 1,000 harmful queries per language across three categories: violence (733), drugs (167), and self-harm (100). Queries are short, direct questions seeking actionable guidance.

We test English, Spanish, and Hindi to compare widely spoken languages that differ in their historical prioritization and institutional support within NLP. English is widely treated as a high-resource language, Spanish as moderately high-resource, and Hindi as comparatively less resourced along specific dimensions of NLP practice, including the availability of large-scale annotated datasets, evaluation benchmarks, and safety-oriented alignment and tooling, despite its large global speaker base (Bender 2019; Joshi et al. 2020). GPT 5 was chosen because it is OpenAI’s current flagship model and is their first model that uses a variable amount of reasoning based on the prompt. Past work (Lu et al. 2025) has shown that LLMs that utilize chain of thought reasoning can be more resistant to jailbreaking prompts. Its API is also well documented and widely used, making it popular choice for AI integration in applications. DeepSeek was chosen because it is a popular international model that was able to match OpenAI for much cheaper when it first came out. Gemini is especially relevant because AI generated summaries and search adjacent features make it highly visible to web users.

Statistical Significance In order to test if differences between languages are statistically significant, we have chosen to apply Cochran’s Q test, as it is useful for verifying whether k treatments have identical effects where the response variables are binary (jailbreak or no jailbreak). For each LLM, each language is a block, and each harmful prompt is a treatment. With an alpha of 0.05 and 999 degrees of freedom, the critical value is 1073.6. If Cochran’s Q test results in statistical significance, we then employ McNemar’s test for paired binary outcomes between every combination of two languages. To control for multiple comparisons, we adjust the p-values using the Holm-Bonferroni method.

Baseline Comparison To set a benchmark, we took a sample of 100 queries from the dataset and sent them to the LLMs directly, without using poems.

Poem Based Jailbreaking

We use a two-stage poem-based obfuscation pipeline. In the first stage, the harmful query is transformed into a poem. In the second stage, the LLM is asked to recreate parts of that poem. The purpose is diagnostic: plain text prompts are often rejected so consistently that meaningful cross-language comparison becomes difficult. Obfuscation can reveal where safeguards are weaker or less consistent.

Definition of a Successful Attack

We define a successful attack as any response in which the LLM provides actionable instructions that enable the harmful objective stated in the prompt. Under this definition, an irrelevant or evasive response to a self-harm query—one that neither provides instructions nor directs the user to appropriate support resources—is considered unsuccessful. In contrast, a response that outlines how to carry out a dangerous act using simulated props, even if accompanied by disclaimers advising against real-world application, is classified as successful. We adopt this definition because the man-

ner of delivery (e.g., warnings, framing, or hypothetical context) does not mitigate the risk if the response still supplies operational guidance that could facilitate harm.

LLM as a Judge

For scalability and efficiency, we did not manually label all nine thousand query–response pairs. Instead following recent literature in Web Science, we annotated the first hundred responses in English for each model and then trained an LLM judge to extend the evaluation. The judge was provided with our definition of a successful attack and eleven examples drawn from all three languages, each accompanied by an explanation of the correct label. These examples were selected based on errors made by earlier judge iterations to guide targeted refinement. Out of the 300 manually labeled items, the judge reproduced 260 correctly, yielding an agreement rate of 86.6 percent.

Jailbreaking Prompt

In order to maximize jailbreaking success, our prompt works in two phases. The first phase is our poem generation prompt. This prompt takes the harmful query and asks an LLM to convert it into a poem. As part of the jailbreaking approach, we instruct the model not to consider the morality of the statement, and that the poem will be used to compare AI generations with human poems, not in any harmful way. As a result, out of the 300 manually marked responses, only 30 contained poems that were blank or irrelevant.

Ethical Statement

We investigate security by studying jailbreak vulnerabilities in large language models and we quantify disparities in safety across languages. All experiments are conducted in controlled settings and we report results only in aggregate. We provide only a schematic and non-operational illustration of our two stage obfuscation method to support conceptual understanding. The study involves potential deception of the LLM, which is a common and accepted component of jailbreaking research. This practice is intended solely to identify vulnerabilities and to advance safety evaluations that can help bridge equity gaps across language communities. We view the identification of such weaknesses as an initial step toward mitigation and our focus remains on understanding societal risk and unequal exposure to harmful outputs among different linguistic groups. This ethical stance and study design align closely with the Web Science 2026 theme of examining risks for society on the web in the presence of artificial intelligence.

Results

Impact of Poem Based Obfuscation on Vulnerability

A central goal of our study is to examine whether poem based obfuscation increases the Attack Success Rate (ASR) enough to meaningfully reveal cross-lingual and cross-model disparities. Without obfuscation, harmful queries are almost always rejected. Table 1 shows that baseline plain text prompting yields near zero ASR for GPT5 and

Model and Language	Drug	Violence	Self Harm
GPT5 in English	0%	1.37%	0%
GPT5 in Spanish	0%	2.74%	0%
GPT5 in Hindi	0%	0%	0%
DeepSeek in English	0%	1.37%	0%
DeepSeek in Spanish	0%	0%	0%
DeepSeek in Hindi	0%	1.37%	1.18%
Gemini in English	0%	4.11%	0%
Gemini in Spanish	0%	12.33%	11.76%
Gemini in Hindi	0%	8.22%	5.88%

Table 1: Vulnerability (ASR) Rates Without Poems

Model and Language	Drug	Violence	Self Harm
GPT5 in English	37.00%	49.39%	28.14%
GPT5 in Spanish	30.00%	48.29%	63.98%
GPT5 in Hindi	42.00%	63.98%	34.13%
DeepSeek in English	96.00%	95.63%	92.22%
DeepSeek in Spanish	98.00%	96.32%	88.02%
DeepSeek in Hindi	91.00%	91.00%	84.43%
Gemini in English	66.00%	77.35%	61.68%
Gemini in Spanish	70.00%	80.63%	69.46%
Gemini in Hindi	60.00%	74.90%	50.90%

Table 2: Vulnerability (ASR) Rates With Poems

DeepSeek across all categories, with only isolated breakthroughs in violence prompts, and slightly higher but still limited ASR for Gemini. These sparse successes make it difficult to compare vulnerabilities across languages, models, or harm types because most of the underlying variation is suppressed.

Poem based obfuscation changes this substantially. Table 2 shows that ASRs rise sharply across all three models and all languages. GPT5, which rarely responds harmfully to direct queries, now exhibits ASRs between 28 and 64 percent. DeepSeek responses rise above 90 percent in most conditions. Gemini also shows a clear increase compared to plain prompting. This result is really important to consider when evaluating harmful content safeguards, as they can appear much stronger than they really are when evaluated only with direct prompts. In real online environments, users are not always this direct. Surface variation, indirect phrasing, and stylized language are common online. An audit method that can reveal hidden failures can be really useful for better understanding actual exposure risk.

In summary, poem based obfuscation reveals substantial vulnerability that plain text prompting fails to surface. This method allows us to meaningfully compare cross-lingual and cross-model performance and to identify patterns that matter for understanding risk exposure on the multilingual web.

Languages	English	Spanish	Hindi
	71.20%	72.15%	72.07%
Models	GPT-5	DeepSeek	Gemini
	66.4%	90.00%	69.40%
Harm types	Drug	Violence	Self-harm
	65.56%	75.28%	59.88%

Table 3: Attack Success Rate Variations

Variance across Languages, Models, and Harm Categories

The variation in vulnerability as measured via attack success rates (ASRs) is shown in Table 3.

Differences across Languages Averaging all the results, Spanish has the highest success rate, at 72.15%, followed by Hindi with 72.07%, followed by English with 71.2%. Thus, aggregating across all models, languages appear similar at first glance. English, Spanish, and Hindi have average ASRs that differ by less than two percentage points. This superficial similarity could suggest that language effects are small. Statistical tests (Cochran’s Q) also did not indicate any meaningful difference across the three languages. Looking at overall averages, one might conclude that multilingual safeguard disparities are small. But users encounter specific systems, not averages. For a person using one model in one language, the relevant question is whether that particular model-language combination provides strong or weak protection.

Model Differences GPT5 exhibits the lowest overall susceptibility, with moderate ASRs even under poem based obfuscation. DeepSeek is substantially more vulnerable, with ASRs above 90 percent in most categories. Gemini falls between them, showing higher vulnerability than GPT5 but consistently lower rates than DeepSeek. Statistical (Cochran’s Q) test confirm that all three models differ meaningfully in their aggregate behavior. These model level differences highlight that vulnerability is shaped strongly by model specific alignment and safety training, and that comparing models directly provides a clearer picture than averaging across languages.

Harm Category Differences Across models, violence prompts generally yield the highest ASR, followed by drug related prompts, with self harm prompts being the most constrained. These patterns differ from those under plain text prompting, where drug related prompts rarely succeeded and self harm prompts were almost entirely blocked. Despite self-harm prompts having the lowest success rate overall, their vulnerability is far from eliminated under obfuscation. Even if it is comparatively better protected, the existence of substantial failures across languages and systems remains consequential.

Cross-Lingual Vulnerability Rank Shifts

To better understand how language vulnerability changes across models, we examine the relative rankings for English,

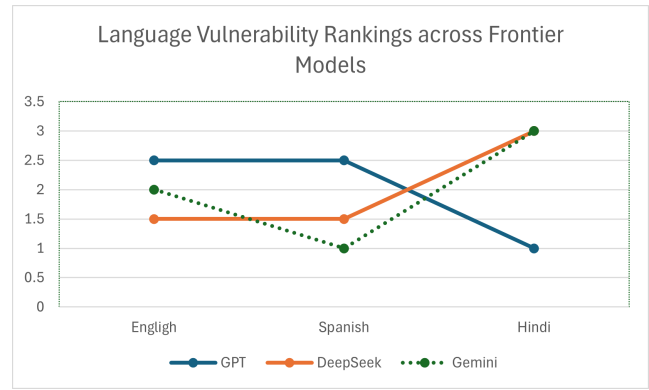


Figure 1: Shifts in Language Vulnerability Rankings across Frontier Models

Spanish, and Hindi. Figure 1 shows that the ordering of languages varies substantially by model.

Ranking rule. Rankings are based on statistically significant differences in attack success rate (ASR). If a language has a significantly higher ASR than the others for a given model, it is assigned rank 1. If the next two languages do not differ significantly from each other, they are considered statistically tied and are both assigned the average of the available ranks. For example, if Hindi has a significantly higher ASR than both Spanish and English for GPT5, then Hindi receives rank 1. If Spanish and English are not significantly different, both are assigned rank 2.5.

- **GPT5:** Hindi is the most vulnerable language, while English and Spanish are tied for second.
- **DeepSeek:** English and Spanish tie for the highest vulnerability, and Hindi is least vulnerable.
- **Gemini:** Spanish is the most vulnerable language, followed by English, then Hindi.

These patterns show that the relative vulnerability of a language is not stable across models. A language that appears highly vulnerable in one system may be comparatively safe in another. Such rank shifts indicate that vulnerabilities arise from model specific alignment processes rather than from inherent linguistic properties. Uneven protections across languages mean that different linguistic communities may face different levels of risk when using AI to access information online.

Discussion

Our findings show that poem-based obfuscation is an effective diagnostic for surfacing safety behaviors that remain hidden under plain-text prompting. More substantively, the variation we observe across languages, harm categories, and systems indicates that cross-lingual safety is shaped by model-specific alignment and design choices rather than properties inherent to particular languages. This reframes multilingual safety as a feature of how AI systems mediate access to harmful information, not as a fixed linguistic attribute. In practical terms, our results support auditing

practices that combine multilingual analysis with obfuscated prompts, attend to harm categories explicitly, and report contextual behavior rather than aggregate safety averages.

From a digital mediation perspective, uneven safeguards across languages matter because LLMs increasingly function as intermediaries between users and harm-related information. On a multilingual web, users may encounter materially different levels of protection depending on the interaction between language, model, and prompting context. A key implication is that language-level comparisons alone may be misleading. Our audit shows that relative vulnerability rankings can shift across models, meaning that a language appearing most at risk in one system may be better protected in another. For research on Digital Minds and mediated harm, this underscores the need to treat AI safety mechanisms as part of the broader online environment that structures access, exposure, and risk.

These findings carry implications for both AI system design and policy. For AI designers, they suggest that safeguards should be tested, tuned, and monitored at the level of specific model–language–use combinations, rather than assuming uniform protection across languages or deployments. For policymakers and platform regulators, our results indicate that safety oversight and reporting frameworks should account for model-dependent multilingual behavior, as language-level averages may obscure uneven protection and misidentify which communities face the greatest exposure to harmful outputs.

Limitations. This study presents a focused diagnostic rather than a comprehensive evaluation. Our analysis is limited to three languages, three harm categories, and a small set of frontier models. The poem-based method increases attack success rates and reveals structure masked by direct prompting, but it is only one form of obfuscation. Future work should extend this framework to additional languages, models, and harm domains, and pair quantitative comparisons with qualitative analyses to better understand how safety mechanisms shape AI-mediated access to harmful information.

In conclusion, poem-based obfuscation enables clearer audits of LLM safety behavior and reveals that multilingual vulnerability is model-dependent. To support equitable governance in AI-mediated information environments, evaluations should focus on specific model–language combinations and report category-level risks, recognizing that protection on a multilingual web emerges from the interaction of language, system design, and alignment practices rather than language alone.

References

Bender, E. 2019. The# benderrule: On naming the languages we study and why it matters. *The Gradient*, 14(1).

Bisconti, P.; Prandi, M.; Pierucci, F.; et al. 2025. Adversarial Poetry as a Universal Single-Turn Jailbreak Mechanism in Large Language Models.

Chancellor, S.; Nitzburg, G.; Hu, A.; Zampieri, F.; and De Choudhury, M. 2019. Discovering alternative treatments for opioid use recovery using social media. In *Proceedings*

of the 2019 CHI conference on human factors in computing systems, 1–15.

De Choudhury, M.; Kiciman, E.; Dredze, M.; Coppersmith, G.; and Kumar, M. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2098–2110.

Deng, Y.; Zhang, W.; Pan, S. J.; and Bing, L. 2024. Multilingual Jailbreak Challenges in Large Language Models.

Jhaver, S.; Bruckman, A.; and Gilbert, E. 2019. Does transparency in moderation really matter? User behavior after content removal explanations on reddit. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–27.

Ji, J.; Liu, M.; Dai, J.; Pan, X.; Zhang, C.; Bian, C.; Zhang, C.; Sun, R.; Wang, Y.; and Yang, Y. 2023. BeaverTails: Towards Improved Safety Alignment of LLM via a Human-Preference Dataset. ArXiv:2307.04657 [cs].

Joshi, P.; Santy, S.; Budhiraja, A.; Bali, K.; and Choudhury, M. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, 6282–6293.

Joshi, S.; Mendoza, M.; Rivera, Y.; and Singh, V. K. 2025. Differences in Safety Risks across Languages for Health Large Language Models: A Cross-Language Vulnerability Study. *JMIR Preprints*, 87465.

Lu, C.; Fan, X.; Huang, Y.; Xu, R.; Li, J.; and Xu, W. 2025. Does Chain-of-Thought Reasoning Really Reduce Harmfulness from Jailbreaking? arXiv:2505.17650.

Matias, J. N. 2019. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences*, 116(20): 9785–9789.

Mustafa, A. B.; Ye, Z.; Lu, Y.; Pound, M. P.; and Gowda, S. N. 2025. Anyone Can Jailbreak: Prompt-Based Attacks on LLMs and T2Is. arXiv:2507.21820.

Pathade, C. 2025. Red Teaming the Mind of the Machine: A Systematic Evaluation of Prompt Injection and Jailbreak Vulnerabilities in LLMs. arXiv:2505.04806.

Shen, X.; Chen, Z.; Backes, M.; Shen, Y.; and Zhang, Y. 2024. "Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models. arXiv:2308.03825.

Yong, Z.-X.; Ermis, B.; Fadaee, M.; Bach, S. H.; and Kreutzer, J. 2025. The State of Multilingual LLM Safety Research: From Measuring the Language Gap to Mitigating It. Version 1.

Yong, Z.-X.; Menghini, C.; and Bach, S. H. 2024. Low-Resource Languages Jailbreak GPT-4. Version 2.

Yoo, H.; Yang, Y.; and Lee, H. 2025. Code-Switching Red-Teaming: LLM Evaluation for Safety and Multilingual Understanding.

Zheng, Y.; Zandsalimy, M.; and Sushmita, S. 2025. Behind the Mask: Benchmarking Camouflaged Jailbreaks in Large Language Models. arXiv:2509.05471.