



iJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IX **Month of publication:** September 2026

DOI: <https://doi.org/10.22214/ijraset.2026.84854>

www.ijraset.com

Call: ☎ 08813907089

E-mail ID: ijraset@gmail.com

A Comprehensive Survey on Code-Switched LLM Safety and Robustness Evaluation

Pulagam Naveen Kumar¹, Sowjanya Bojja², Banala Ramyasree³, Kaleru Anoosha⁴

^{1, 2, 3}Assistant Professor, Department of Computer Science & Engineering, Scient Institute of Technology, Ibrahimpatnam, R.R Dist [India]

⁴Assistant Professor, Department of AIML, Scient Institute of Technology, Ibrahimpatnam, R.R Dist [India]

Abstract: Large language models (LLMs) have demonstrated remarkable capabilities across languages, yet their safety alignment remains predominantly evaluated in monolingual, especially English, settings. Code-switching (also known as code-mixing)—the alternation between two or more languages within a single utterance or conversation—is a pervasive phenomenon in multilingual societies and digital communication. Recent evidence reveals that code-switched inputs can substantially degrade the safety robustness of LLMs, enabling jailbreaks and harmful outputs that monolingual safety mechanisms fail to intercept. This survey provides the first comprehensive review of research on code-switched LLM safety and robustness evaluation. We systematize existing attack methodologies (including Code-Switching Red-Teaming, Multilingual Blending, and attributional analyses), evaluation datasets and metrics, empirical findings on attack success rates and linguistic factors, and emerging mitigation strategies. We further situate these works within the broader multilingual safety literature, highlight critical gaps in linguistic coverage, cultural contextualization, and mechanistic understanding, and outline a research agenda toward equitable, linguistically inclusive LLM safety. Our synthesis aims to guide researchers and practitioners in developing more robust evaluation frameworks and alignment techniques for real-world multilingual deployments.

Keywords: Large Language Models, Code-Switching, Code-Mixing, Safety Evaluation, Robustness, Red-Teaming, Jailbreak, Multilingual NLP, Alignment

I. INTRODUCTION

The rapid advancement and widespread deployment of large language models (LLMs) have elevated safety and alignment to central concerns in artificial intelligence research and practice. Safety evaluation typically probes models for toxic, biased, harmful, or policy-violating outputs under adversarial or naturalistic inputs. However, the overwhelming majority of existing safety benchmarks, red-teaming protocols, and alignment datasets remain English-centric or strictly monolingual. This creates a critical mismatch with real-world usage patterns: billions of speakers routinely engage in code-switching (CSW) or code-mixing—the fluid alternation of languages, scripts, or dialects within a single discourse unit.

Code-switching is not merely a linguistic curiosity; it is a dominant mode of communication in multilingual communities across South Asia, Africa, the Middle East, Latin America, and diaspora populations worldwide. In digital spaces—social media, messaging applications, and online forums—code-mixed text is ubiquitous. Recent studies have demonstrated that LLMs, despite strong monolingual safety performance, exhibit markedly higher rates of harmful compliance when prompts incorporate intra-sentence or inter-sentence code-switching. Attack success rates (ASR) can rise by tens of percentage points, and in some configurations exceed 60–90%, even for models that refuse the equivalent English request.

These findings expose a fundamental robustness gap: current safety alignment techniques appear brittle under linguistic forms that deviate from the high-resource monolingual data on which they were primarily trained and evaluated. The phenomenon raises urgent questions for both research and deployment. How should safety evaluation be redesigned to account for code-switching? What linguistic and cultural factors modulate vulnerability? Can mechanistic insights (e.g., attention or saliency drift) explain the failures? And what mitigation strategies are effective and scalable?

This survey addresses these questions by providing a systematic synthesis of the emerging literature on code-switched LLM safety and robustness. Our contributions are as follows:

- 1) We formalize the problem space and clarify terminology (code-switching vs. code-mixing, intra- vs. inter-sentence, script vs. transliteration).
- 2) We review and taxonomize existing evaluation frameworks, datasets, and attack methods, with particular attention to Code-Switching Red-Teaming (CSRT), Multilingual Blending, and attributional analyses.

- 3) We synthesize empirical findings on attack effectiveness, the role of language resource levels, morphological and typological factors, and cultural context.
- 4) We catalog proposed mitigations and assess their strengths and limitations.
- 5) We identify open challenges and propose a concrete research agenda for the community.

The remainder of the paper is organized as follows. Section 2 provides background on code-switching linguistics and multilingual LLM capabilities. Section 3 reviews the broader multilingual safety landscape. Section 4 presents a detailed survey of code-switched safety evaluation methods and findings. Section 5 discusses mitigations. Section 6 articulates open challenges and future directions. Section 7 concludes.

II. BACKGROUND

A. Code-Switching and Code-Mixing

Code-switching (CSW) refers to the practice of alternating between two or more languages or language varieties within a single conversation or utterance. Linguists distinguish several dimensions: (i) intra-sentential vs. inter-sentential switching; (ii) insertional vs. alternational patterns; and (iii) the role of a matrix language versus embedded language. In computational literature, the term “code-mixing” is often used interchangeably or to emphasize intra-sentential mixing that produces hybrid lexical and syntactic forms. A further important distinction arises between script-preserving code-switching (each language retains its orthography) and transliterated or romanized forms (e.g., Roman Urdu, Hinglish, Arabizi), which are extremely common in informal digital communication.

From an NLP perspective, code-switched text poses challenges for tokenization, language identification, syntactic parsing, and semantic understanding. Pre-LLM research established that models trained exclusively on monolingual data systematically underperform on CSW inputs. Even modern multilingual LLMs, despite exposure to some mixed-language data during pre-training, frequently exhibit degraded grammaticality, factual consistency, and—critically—safety behavior under code-switched prompts.

B. Multilingual Capabilities of LLMs

Contemporary LLMs are trained on massive multilingual corpora and support dozens to hundreds of languages. Performance, however, remains highly uneven: high-resource languages (English, Chinese, Spanish, etc.) benefit from denser representation, while mid- and low-resource languages lag in both capability and safety alignment. Cross-lingual transfer of alignment is imperfect; safety behaviors learned primarily in English do not reliably generalize to other languages, let alone to mixed-language inputs. This asymmetry is a recurring theme in the safety literature reviewed below.

C. Safety Evaluation Concepts

Key concepts in LLM safety evaluation include: red-teaming (adversarial probing to elicit undesirable behaviors), jailbreaking (bypassing safety guardrails), attack success rate (ASR; the fraction of prompts that elicit harmful responses), refusal rate, and various harm taxonomies (toxicity, bias, violence, criminal activity, self-harm, etc.). Evaluation may be automatic (classifier-based) or human-judged. Robustness refers to the stability of safe behavior under distributional shift, including linguistic perturbations such as code-switching.

III. THE MULTILINGUAL LLM SAFETY LANDSCAPE

Before focusing on code-switching specifically, it is essential to situate the problem within the broader multilingual safety literature. A systematic review of nearly 300 LLM safety publications from major *ACL venues (2020–2024) reveals a pronounced English-centric bias: the vast majority of safety research targets English, with even high-resource non-English languages receiving limited dedicated attention. Non-English languages are frequently included only as secondary evaluation languages in multilingual benchmarks rather than as primary objects of study. Language documentation practices in English-centric papers are often incomplete.

This language gap has direct consequences for code-switched evaluation. Because safety alignment data and evaluation protocols are overwhelmingly monolingual (and English-heavy), models learn safety associations that are tightly coupled to particular lexical and syntactic forms. When those forms are disrupted by language mixing, the associations weaken—a phenomenon analyzed in detail by attributional studies. The multilingual safety literature further documents that low-resource languages and non-standard orthographies (e.g., Arabizi) can themselves serve as jailbreak vectors, reinforcing the broader pattern that deviation from high-resource monolingual English increases vulnerability.

Several surveys and position papers have called for more inclusive multilingual safety research, recommending culturally grounded benchmarks, diverse training data generation, and deeper investigation of cross-lingual safety generalization. Code-switching constitutes one of the most naturalistic and challenging test cases for these recommendations.

IV. CODE-SWITCHED SAFETY AND ROBUSTNESS EVALUATION

A. Code-Switching Red-Teaming (CSRT)

The seminal work introducing systematic code-switching red-teaming is Yoo, Yang, and Lee (2024/2025). They propose CSRT, a fully automated framework that synthesizes code-switching red-teaming queries by combining translations of existing adversarial prompts across up to ten languages (spanning high-, mid-, and low-resource tiers: English, Chinese, Italian, Vietnamese, Arabic, Korean, Thai, Bengali, Swahili, Javanese). The resulting CSRT dataset comprises 315 queries designed to elicit a wide range of undesirable behaviors.

Extensive experiments on ten state-of-the-art LLMs (both open and proprietary) demonstrate that CSRT substantially outperforms prior multilingual red-teaming techniques such as MultiJail and Sandwich Attack. Notably, CSRT achieves approximately 46.7% higher attack success rates than standard English attacks and proves effective across conventional safety categories (hate speech, bias, undesired information, violent and non-violent crime). Ablation studies reveal that attack success increases with the number of languages involved and is particularly amplified when low-resource languages are included. The authors further document an unintended correlation between language resource availability and the strength of safety alignment, and they show that CSRT can be extended using monolingual seed data.

Beyond attack success, CSRT simultaneously probes multilingual understanding: models must both comprehend the mixed-language query and decide whether to comply with the harmful intent. This dual evaluation is a distinctive strength of the approach.

B. Multilingual Blending

Song et al. (2024/2025) introduce Multilingual Blending, a mixed-language query-response scheme that evaluates safety alignment under sophisticated multilingual conditions without requiring hand-crafted adversarial templates. By constructing prompts that interleave multiple languages, they observe dramatically elevated bypass rates: 67.23% on GPT-3.5 and 40.34% on GPT-4o, far exceeding single-language baselines (which often remain near zero for English). Effectiveness varies with linguistic properties: languages of differing morphology and from diverse language families tend to produce stronger attacks. These results underscore that even relatively simple mixing strategies can compromise safeguards that appear robust in monolingual settings.

C. Attributional Analyses of Safety Failures

Banerjee et al. (2025) provide a mechanistic perspective through the lens of attributional collapse. Their systematic evaluation of open models shows ASR rising from approximately 9% in monolingual English to 69% under code-mixed inputs, with rates exceeding 90% in certain non-Western cultural contexts (Arabic, Hindi). Effects persist on both synthetic datasets and real-world social media traces, indicating practical risk for large user populations.

To explain the phenomenon, the authors introduce saliency drift attribution (SDA). Under code-mixing, the model's internal attention and attribution mass drift away from safety-critical tokens (e.g., "violence," "corruption," "destruction"), effectively blinding the safety mechanisms to harmful intent even when semantics are preserved. Targeted lexical perturbations (replacing toxic or non-toxic words) do not reproduce the same vulnerability, suggesting that the distributed, multi-language context is essential to the failure mode. A lightweight translation-based restoration strategy recovers roughly 80% of the lost safety, offering a practical, model-agnostic mitigation path.

D. Additional Empirical Studies

Several complementary studies expand the evidence base. Work on Roman Urdu demonstrates consistent degradation of safety detection under code-switched and transliterated inputs across passive suicidal ideation, psychological distress, threat, and coercion categories; instruction-tuned and reasoning-capable models show relatively greater robustness, suggesting alignment gaps rather than fundamental incapacity. Phonetic code-mixing perturbations (altering spellings while preserving pronunciation) further expose tokenizer-level vulnerabilities. Emerging work on large audio language models indicates that code-switched speech similarly elevates jailbreak success rates, extending the problem beyond text.

Collectively, these studies establish code-switching—whether script-mixed, transliterated, phonetic, or spoken—as a potent and under-addressed attack surface.

E. Datasets, Metrics, and Evaluation Practices

Key resources include the CSRT dataset (315 queries, up to 10 languages), MultiJail and its derivatives, culturally contextualized “Local” versus “Generic” test sets, the Roman Urdu Adversarial Benchmark (RUAB), and various synthetic mixing pipelines. Primary metrics are attack success rate (ASR), refusal rate, comprehension/accuracy of the mixed input, and, in mechanistic studies, saliency or attribution scores. Evaluation remains challenging: automatic harm classifiers trained predominantly on English may themselves be unreliable on code-switched text; human annotation is costly and requires bilingual expertise; and reproducibility is hindered by proprietary model APIs and evolving safety filters.

V. MITIGATION STRATEGIES

Proposed defenses remain relatively preliminary. Translation-based restoration—detecting code-mixed input, translating to a high-resource language (typically English), applying the safety filter, and optionally translating the response back—recovers a substantial fraction of lost safety in controlled experiments.

Other directions include: (i) expanding safety fine-tuning and preference data to include realistic code-switched examples; (ii) multi-lingual or code-switched preference optimization; (iii) improved language-agnostic or cross-lingual safety representations; and (iv) runtime detection of mixed-language harmful intent. No single method currently offers comprehensive protection across languages, scripts, and cultural contexts, and most approaches introduce latency or require additional model capacity.

VI. OPEN CHALLENGES AND FUTURE DIRECTIONS

Despite rapid progress, substantial gaps remain:

- 1) Linguistic and cultural coverage. Existing benchmarks cover a limited set of language pairs and rarely incorporate deep cultural knowledge or locally sensitive harm categories. Low-resource and indigenous languages are almost entirely absent.
- 2) Mechanistic understanding. While saliency drift provides an initial explanation, finer-grained accounts of how safety circuits interact with multilingual representations are needed. Causal interventions and circuit-level analyses remain scarce.
- 3) Multi-turn and interactive settings. Most evaluations are single-turn. Real conversations involve multi-turn code-switching, topic drift, and social context that may further erode safety.
- 4) Multimodal extension. Code-switched speech, image+text, and video settings are only beginning to be explored.
- 5) Evaluation validity. Automatic metrics and English-centric harm classifiers require rigorous validation on code-switched data. Community standards for reproducible, open evaluation are underdeveloped.
- 6) Alignment data and methods. Scalable generation of high-quality code-switched safety data, and training objectives that explicitly preserve safety salience across linguistic forms, constitute open research problems.

We recommend that future work prioritize: (1) community-driven, culturally grounded CSW safety benchmarks spanning diverse language families and scripts; (2) mechanistic and causal studies of safety failure under mixing; (3) multi-turn and multimodal evaluation protocols; (4) open, reproducible evaluation infrastructure; and (5) alignment techniques that treat code-switching as a first-class phenomenon rather than an afterthought.

VII. CONCLUSION

Code-switching is a fundamental feature of human multilingual communication and an increasingly recognized attack surface for large language models. Empirical evidence from CSRT, Multilingual Blending, attributional analyses, and related studies demonstrates that safety guardrails optimized for monolingual English (and, to a lesser extent, other high-resource languages) are systematically fragile under mixed-language inputs. Attack success rates rise substantially, sometimes dramatically, and the failures appear rooted in attributional and representational mismatches rather than mere incomprehension.

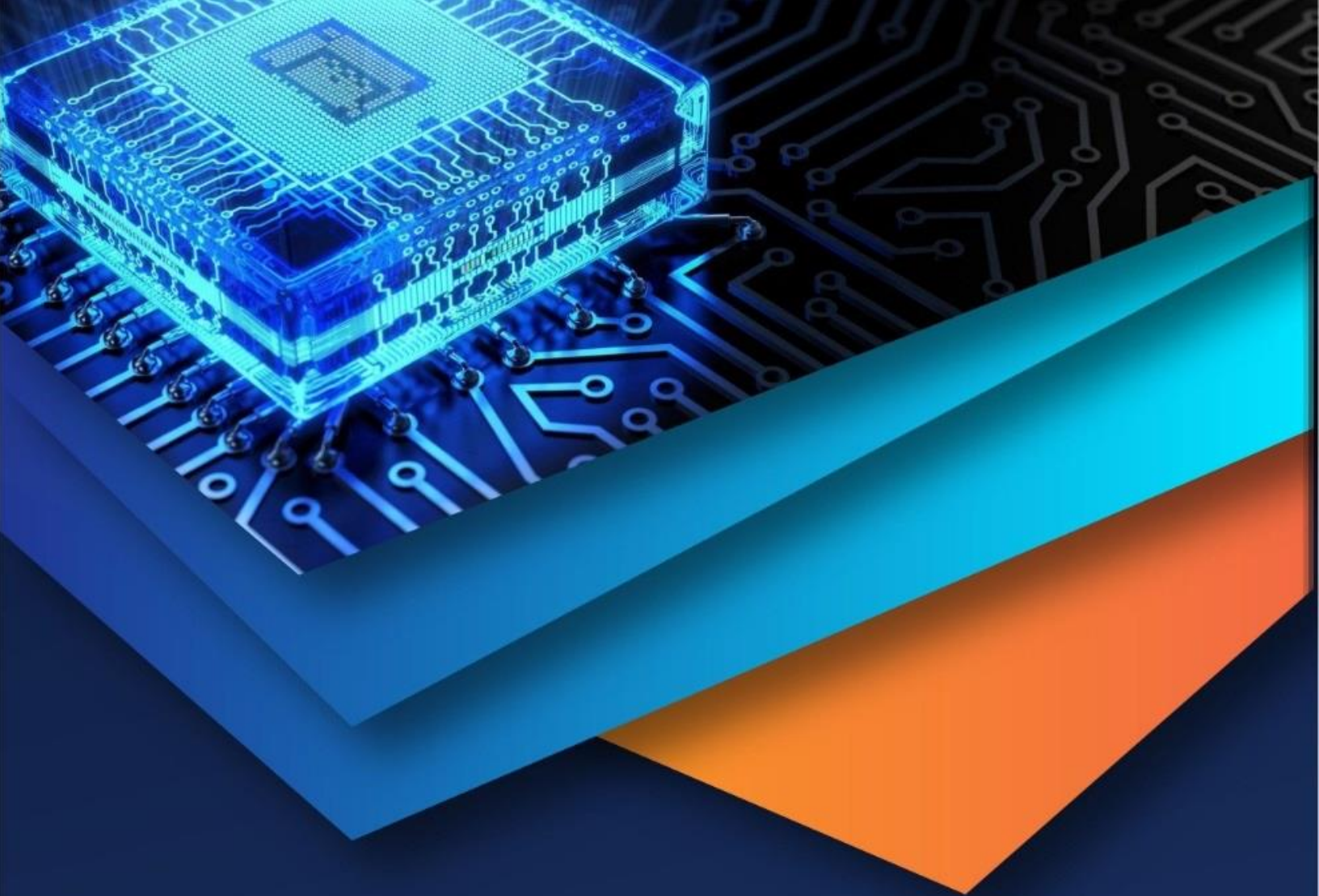
Addressing this vulnerability is both a technical and an equity imperative. Billions of users communicate in code-switched forms; safety systems that fail them are incomplete. By systematizing the current literature, clarifying evaluation practices, and charting open problems, this survey aims to accelerate the development of linguistically inclusive, robust LLM safety evaluation and alignment. The path forward requires collaboration across NLP, linguistics, safety research, and the communities whose languages and communicative practices have been historically underrepresented in AI development.

REFERENCES

- [1] Banerjee, S., Chatterjee, P., Kumar, S., Layek, S., Agrawal, P., Hazra, R., & Mukherjee, A. (2025). Attributional Safety Failures in Large Language Models under Code-Mixed Perturbations. arXiv preprint arXiv:2505.14469.



- [2] Gupta, H., Jayarao, P., Varshney, N., & Dwivedi, C. (2026). Code-Mixing and Code-Switching for Text in the LLM Era: A Playbook of Models, Data, Evaluation, and Open Problems. arXiv preprint arXiv:2602.11181.
- [3] Sheth, R., Sinha, S. R., Patil, M., Beniwal, H., & Singh, M. (2025). Beyond Monolingual Assumptions: A Survey of Code-Switched NLP in the Era of Large Language Models. arXiv preprint arXiv:2510.07037.
- [4] Song, J., Huang, Y., Zhou, Z., & Ma, L. (2025). Multilingual Blending: Large Language Model Safety Alignment Evaluation with Language Mixture. In Findings of the Association for Computational Linguistics: NAACL 2025, pages 3433–3449.
- [5] Winata, G. I., Aji, A. F., Yong, Z.-X., & Solorio, T. (2023). The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges. In Findings of the Association for Computational Linguistics: ACL 2023, pages 2936–2978.
- [6] Yong, Z.-X., Ermis, B., Fadaee, M., Bach, S. H., & Kreutzer, J. (2025). The State of Multilingual LLM Safety Research: From Measuring the Language Gap to Mitigating It. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 15845–15860.
- [7] Yoo, H., Yang, Y., & Lee, H. (2025). Code-Switching Red-Teaming: LLM Evaluation for Safety and Multilingual Understanding. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13392–13413.



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)