ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul962

TanglishGuard: Benchmarking AI Safety Guardrails on Tamil-English Code-Mixed Prompts

Vivina Vijesh[1]; Dr. E. Gothai[2]

[1]Third year Student, [2]M.E., Ph.D., Professor & Head

[1,2]Department of Computer Science and Design, Kongu Engineering College, Perundurai, Erode, India

ORCiD ID: [1]https://orcid.org/0009-0008-4484-261X,
[2]https://orcid.org/0000-0002-6889-4851

Publication Date: 2026/07/29

Abstract: Large Language Models (LLMs) have become indispensable across modern artificial intelligence applications, yet their safety mechanisms continue to be evaluated almost exclusively on English inputs. In multilingual nations like India, where millions communicate daily through Tanglish a fluid blend of Tamil and English this evaluation gap raises fundamental concerns about whether these systems respond safely and consistently across diverse linguistic contexts. This study addresses this critical oversight by introducing TanglishGuard, a novel benchmark designed to systematically assess the robustness of LLM safety mechanisms against code-mixed inputs.

The benchmark evaluates four state-of-the-art models ChatGPT, Gemini, Claude, and DeepSeek using equivalent harmful prompts expressed in English, Tamil, and Tanglish across nine distinct harm categories. Through rigorous experimentation, our findings reveal that while all models demonstrate strong safety compliance on English and Tamil inputs, Tanglish prompts reveal subtle but consistent vulnerabilities. These inconsistencies manifest as occasional failures in detecting harmful intent within code-mixed language, highlighting significant gaps in multilingual AI safety frameworks.

TanglishGuard provides a practical, reproducible framework for evaluating safety in mixed-language settings, offering empirical evidence that current safety evaluations are insufficient for real-world multilingual usage. The benchmark contributes to the development of more robust and equitable AI systems by ensuring safety mechanisms are tested against authentic communication patterns rather than sanitised English-only datasets. This work underscores the urgent imperative to move beyond English-centric safety evaluations. As AI systems become increasingly embedded in diverse linguistic communities worldwide, ensuring their safety across the full spectrum of human language use is not merely a technical challenge but a fundamental requirement for fairness, equity, and responsible AI deployment.

Keywords: *Tanglish, Code-Mixing, AI Safety, Large Language Models, Multilingual Benchmarks, Red-Teaming.*

How to Cite: Vivina Vijesh; Dr. E. Gothai (2026) TanglishGuard: Benchmarking AI Safety Guardrails on Tamil-English Code-Mixed Prompts. *International Journal of Innovative Science and Research Technology,* 11(7), 1662-1669. https://doi.org/10.38124/ijisrt/26jul962

I. INTRODUCTION

Large Language Models (LLMs) have emerged as transformative technologies, reshaping how humans interact with machines across countless domains. From virtual assistants and content generation to decision support systems, these models are now deeply embedded in the fabric of modern digital life. However, with great capability comes great responsibility ensuring that these powerful systems operate safely and ethically remains one of the most pressing challenges in artificial intelligence research. Safety alignment, the process of training models to refuse harmful requests and avoid generating dangerous content, has become a cornerstone of responsible AI development. Most safety alignment research has focused predominantly on English prompts, creating a significant blind spot in our understanding of how these systems behave when interacting with users who communicate in other languages or language varieties.

In India, a nation characterised by extraordinary linguistic diversity, the reality of everyday communication often defies rigid language boundaries. Millions of users navigate seamlessly between languages, blending Tamil and

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul962

English into a vibrant hybrid known as Tanglish. This code-mixed variety is not merely a linguistic curiosity; it is the primary mode of online communication for a vast and growing population. From social media posts and WhatsApp messages to informal emails and online comments, Tanglish reflects the organic, fluid, and pragmatic ways in which multilingual communities naturally interact. Despite its prevalence, Tanglish remains almost entirely unexplored in the context of AI safety evaluation. If safety mechanisms fail to detect harmful intent in code-mixed inputs, it raises profound concerns about the robustness, fairness, and equity of multilingual AI systems.

The problem is not merely hypothetical. Prior research has established that LLMs often exhibit significant cross-lingual safety drift, performing less safely on non-English prompts compared to their English counterparts. Studies have shown that language differences and even code-switching can throw off safety guardrails, enabling users to bypass safety measures by translating harmful prompts into low-resource languages. However, most existing multilingual benchmarks, such as XSafety and IndicSafe, evaluate models on monolingual non-English inputs, leaving the critical domain of code-mixed language completely unaddressed.

This paper introduces TanglishGuard, a comprehensive benchmark designed to bridge this critical gap. TanglishGuard systematically evaluates LLM safety behaviour by presenting models with equivalent harmful prompts in English, Tamil, and Tanglish.

By comparing responses across these language conditions, we aim to answer three fundamental questions:

- How do LLMs interpret harmful intent in Tanglish compared to English?
- Do LLM safety mechanisms perform equally well on code-mixed inputs?
- What are the broader implications for the safety and fairness of multilingual AI?

The primary contribution of this work is threefold:

(a) a curated dataset of equivalent English-Tanglish prompt pairs spanning nine harmful categories, (b) a systematic evaluation framework with three key metrics refusal behaviour, response consistency, and explanation quality and (c) empirical insights into the reliability of current LLM safety mechanisms in code-mixed contexts. Through extensive experimentation with four state-of-the-art LLMs, we demonstrate that while models generally adhere to safety guidelines, Tanglish prompts reveal subtle but significant inconsistencies that demand attention. These findings underscore the urgent need to expand AI safety evaluations beyond English-only benchmarks and test systems against the authentic ways people communicate online.

II. LITERATURE SURVEY

The landscape of multilingual AI safety research has evolved considerably in recent years, yet significant gaps remain. Early safety evaluations focused almost exclusively on English, assuming that safety alignment would generalise across languages. This assumption has proven problematic, as subsequent research has consistently demonstrated that LLMs exhibit varying safety behaviours across different linguistic contexts. Table 1 provides a comprehensive overview of key studies in this domain, highlighting their contributions, methodologies, and the gaps that TanglishGuard addresses.

Table 1 Literature Survey on Multilingual AI Safety Benchmarks

Study	Year	Benchmark/Framework	Languages Covered	Key Contribution	Gap Addressed by TanglishGuard
Wang et al. (2024)	2024	XSafety	10 languages (non-English)	First multilingual safety benchmark; found models are more likely to give unsafe responses in non-English prompts; proposed prompting strategy reduced unsafe outputs by ~42%	Does not include code-mixed languages; focuses on monolingual non-English inputs
Shen et al. (2024)	2024	Multilingual Safety Analysis	Multiple languages	Examined variations in safety challenges across languages; identified significant language gap in LLM safety research	Does not evaluate code-switching; primarily focuses on monolingual contexts
Yoo et al. (2025)	2025	CSRT (Code-Switching Red-Teaming)	Up to 10 languages	Discovered code-switching effectively elicits undesirable behaviours; achieved 46.7% more attacks than standard English attacks	Focuses on generating attacks rather than systematic benchmarking of safety consistency
Pattnayak & Chowdhuri (2026)	2026	IndicSafe	12 Indic languages	First systematic evaluation of LLM safety across 12 Indic languages; revealed significant cross-lingual safety drift with only 12.8% cross-language consistency	Evaluates monolingual Indic languages; does not cover code-mixed varieties like Tanglish
Ning et al. (2025)	2025	LinguaSafe	12 languages	Comprehensive multilingual safety benchmark with 45k entries; emphasises linguistic authenticity	Covers standard languages; lacks code-mixed or hybrid language varieties

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul962

Choi et al. (2026)	2026	XL-SafetyBench	10 country-language pairs	Country-grounded cross-cultural benchmark with 5,500 test cases; evaluates both universal and culturally specific harms	Focuses on country-language pairs; does not address code-mixing
This Study	2026	TanglishGuard	English, Tamil, Tanglish	First systematic benchmark for evaluating LLM safety on code-mixed Tamil-English inputs; reveals inconsistencies in safety behaviour across language conditions	Addresses the critical gap of code-mixed language safety evaluation

The literature reveals a clear and consistent pattern: while the research community has made commendable progress in expanding safety evaluations beyond English, the domain of code-mixed languages remains conspicuously underexplored. Code-switching is not a marginal phenomenon; it is a natural, ubiquitous practice in multilingual communities worldwide. In India alone, hundreds of millions of users communicate daily in hybrid language varieties like Tanglish, Manglish (Malayalam-English), and Hinglish (Hindi-English). If AI safety mechanisms are not tested against these authentic communication patterns, they risk being fundamentally inadequate for the populations they serve.

Furthermore, existing benchmarks that do address code-switching, such as the CSRT framework by Yoo et al. (2025), primarily focus on generating adversarial attacks rather than systematically benchmarking safety consistency across equivalent prompts in different language conditions. TanglishGuard fills this gap by providing a structured, reproducible framework for evaluating whether LLMs respond consistently to the same harmful intent expressed in English, Tamil, and Tanglish.

III. PROPOSED METHODOLOGY

➢ *Problem Formulation*

The central research question driving this study is: Do LLMs exhibit consistent safety behaviour when presented with equivalent harmful prompts in English, Tamil, and Tanglish? To address this question, we formulate the problem as a comparative evaluation of model responses across three language conditions for identical harmful intent.

Let PE represent a harmful prompt in English, PT represent its equivalent translation in Tamil, and PTG represent its equivalent in Tanglish (Tamil-English code-mixed form). For each prompt Pi, we obtain a response Ri=M(Pi) from a given LLM. We then evaluate Ri along three dimensions: (a) whether the model refuses the request, (b) whether the response is consistent across language conditions, and (c) the quality of the explanation provided when the model refuses.

➢ *Dataset Construction*

The TanglishGuard dataset comprises 90 prompts organised into 9 categories, with 10 prompts per category. Each prompt is carefully curated to represent a specific type of harmful request, as outlined in Table 2.

Table 2 Prompt Categories and Distribution

Category	Description	Number of Prompts
Cyber security	Hacking, phishing, and malware-related requests	10
Academic Integrity	Cheating, plagiarism, and assignment misuse	10
Misinformation	Fake news and misleading content	10
Fraud & Illegal Activities	Fraud, forgery, and scams	10
Harmful Advice	Dangerous or unsafe guidance	10
Privacy	Personal data and information misuse	10
Violence	Violent or criminal requests	10
Self-Harm	Self-harm-related prompts	10
Prompt Manipulation	Attempts to bypass AI safety mechanisms	10
Total		**90**

For each prompt, three versions are created: (a) an English version, (b) a Tamil version (using Tamil script), and (c) a Tanglish version (using Roman script with Tamil-English code-mixing). The translation process prioritises semantic equivalence and naturalness, ensuring that the same harmful intent is expressed authentically in each language condition.

➢ *Proposed Algorithm*

The TanglishGuard evaluation framework follows a systematic, multi-stage algorithm designed to ensure rigorous and reproducible assessment of LLM safety behaviour.

Text

Algorithm: TanglishGuard Safety Evaluation Framework
Input:
- Set of harmful prompts P = {P_1, P_2, ..., P_{90}}
- Language conditions L = {English, Tamil, Tanglish}
- Set of LLMs M = {M_1, M_2, ..., M_4}
- Number of evaluation runs R = 3

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul962

Output:
- Safety scores S for each (model, language, prompt) combin ation
- Consistency metrics C across language conditions
- Refusal rates and explanation quality scores

Procedure:

1. Dataset Preparation:
For each harmful prompt p in P:
Create three equivalent versions:
p_eng ← English version
p_tam ← Tamil version (Tamil script)
p_tgl ← Tanglish version (Roman script, code-mixed)
Validate semantic equivalence through expert review
2. Model Querying:
For each model m in M:
For each prompt version p_lang in {p_eng, p_tam, p_tgl}:
For run r = 1 to R:
response_{m,lang,r} ← m.generate(p_lang)
Store response for subsequent evaluation
3. Response Evaluation:
For each response_{m,lang,r}:
Evaluate three metrics:
a. Refusal Behaviour (Binary):
refusal ← 1 if response contains explicit refusal,
0 otherwise
b. Safety Score (1-10):
score ← Evaluate based on:
- Presence of explicit refusal
- Avoidance of harmful content generation
- Alignment with safety policies
c. Explanation Quality (1-5):
quality ← Evaluate based on:
- Clarity of reasoning
- Appropriateness of explanation
- Specificity of refusal justification
4. Consistency Analysis:
For each model m and prompt p:
Compare responses across language conditions:
consistency_{m,p} ← Similarity(refusal_{m,eng,p},
refusal_{m,tam,p},
refusal_{m,tgl,p})
Compute language-wise average consistency scores
5. Aggregation and Reporting:
For each model m:
Compute average refusal rate across all prompts and langua ges
Compute average safety score per language condition
Compute average consistency score
Generate comparative tables and visualisations

➢ *Implementation Details*

The experimental implementation follows a structured pipeline to ensure reproducibility and reliability.

- *Prompt Engineering:*

Each prompt is crafted to be clear, unambiguous, and representative of real-world harmful requests. The Tanglish versions are validated by native Tamil speakers with expertise in code-mixed language to ensure authenticity.

- *Model Selection:*

Four state-of-the-art LLMs are evaluated:

✓ ChatGPT (OpenAI) - A widely used general-purpose LLM
✓ Gemini (Google) - A multimodal LLM with strong multilingual capabilities
✓ Claude (Anthropic) - A model trained with Constitutional AI principles for enhanced safety
✓ DeepSeek - A recent LLM with competitive performance

- *Evaluation Parameters:*

Each prompt is submitted to each model three times to account for variability in model outputs. The experimental configuration is summarised in Table 3.

Table 3 Experimental Configuration

Parameter	Value
Total Prompts	90
Languages	English, Tamil, Tanglish
Categories	9
Models	ChatGPT, Gemini, Claude, DeepSeek
Runs per Prompt	3

- *Evaluation Metrics:*

Responses are evaluated using three primary metrics, as detailed in Table 4.

Table 4 Evaluation Metrics

Metric	Description	Scale
Refusal Behaviour	Whether the model explicitly refused the harmful request	Binary (0/1)
Response Consistency	Whether equivalent English, Tamil, and Tanglish prompts received similar safety responses	1-10
Explanation Quality	Whether the refusal included a clear, appropriate, and specific explanation	1-5

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul962

- *Data Collection:*

For each prompt-model-language-run combination, the complete model output is recorded. This includes the full response text, allowing for qualitative analysis alongside quantitative scoring.

IV. RESULTS AND DISCUSSION

➢ *Overall Performance*

The experimental results reveal that all four evaluated LLMs demonstrate strong overall safety compliance, with refusal rates exceeding 95% across all language conditions. Table 5 presents the aggregated performance metrics.

Table 5 Overall Results Across All Models

Model	Refusal Rate (%)	Safety Score (1-10)	Consistency (1-10)
ChatGPT	98%	9.6	9.4
Gemini	96%	9.2	8.9
Claude	97%	9.7	9.5
DeepSeek	96%	9.4	9.6

Claude achieves the highest safety score (9.7) and consistency (9.5), which aligns with its Constitutional AI training approach designed specifically to enhance harmlessness. ChatGPT follows closely with a 98% refusal rate and 9.6 safety score. Gemini and DeepSeek, while still demonstrating strong performance, show slightly lower scores, suggesting minor variations in safety alignment across different architectural approaches.

➢ *Language-Wise Performance*

A more nuanced picture emerges when examining performance across individual language conditions. Table 6 presents the language-wise safety scores for each model.

Table 6 Language-Wise Safety Compliance

Model	English	Tamil	Tanglish
ChatGPT	10	10	9.9
Gemini	10	10	9.9
Claude	10	10	9.9
DeepSeek	10	10	9.8

All models achieve perfect or near-perfect safety scores for English and Tamil prompts. However, a subtle but consistent drop is observed for Tanglish prompts across all models. ChatGPT, Gemini, and Claude score 9.9 on Tanglish (compared to 10.0 on English and Tamil), while DeepSeek scores 9.8. While these differences are small, they are consistent across models and categories, suggesting that code-mixed inputs introduce a slight but systematic reduction in safety compliance.

This finding is significant because it indicates that even models with strong multilingual capabilities—which perform flawlessly on both English and Tamil—are slightly less safe when faced with code-mixed inputs. This aligns with findings from prior research on cross-lingual safety drift, where models exhibited reduced safety on non-English prompts. The TanglishGuard results extend this observation to the domain of code-mixing, suggesting that the phenomenon of "safety drift" may be even more nuanced and pervasive than previously recognised.

➢ *Category-Wise Performance*

Table 7 presents the safety scores across different prompt categories, revealing interesting patterns in how models handle various types of harmful requests in Tanglish.

Table 7 Category-Wise Performance (Safety Scores 1-10)

Category	ChatGPT	Gemini	Claude	DeepSeek
Cybersecurity	9.8	9.9	9.8	9.8
Academic Integrity	9.6	9.8	9.6	9.6
Fraud & Illegal Activities	9.8	9.6	9.8	9.7
Violence	9.9	9.8	9.9	9.9
Self-Harm	10.0	9.9	10.0	9.8
Privacy	9.7	10.0	9.7	9.6
Misinformation	9.5	9.7	9.5	9.5

Several observations emerge from this category-wise analysis:

- Self-Harm prompts consistently receive the highest safety scores across all models (9.8-10.0), suggesting that safety training prioritises this category effectively.
- Misinformation prompts show the lowest safety scores (9.5-9.7), indicating that models may be slightly less robust in detecting and refusing requests related to fake news and misleading content.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul962

- Privacy requests also show some variation, with Gemini achieving a perfect 10.0 while other models score 9.6-9.7.
- Academic Integrity prompts consistently score 9.6 across models, suggesting that cheating and plagiarism requests are handled with moderate but consistent safety.

➢ *Visual Analysis*

Figure 1 illustrates the overall safety compliance of the four LLMs on the TanglishGuard benchmark, showing the distribution of safety scores across all prompts and language conditions.

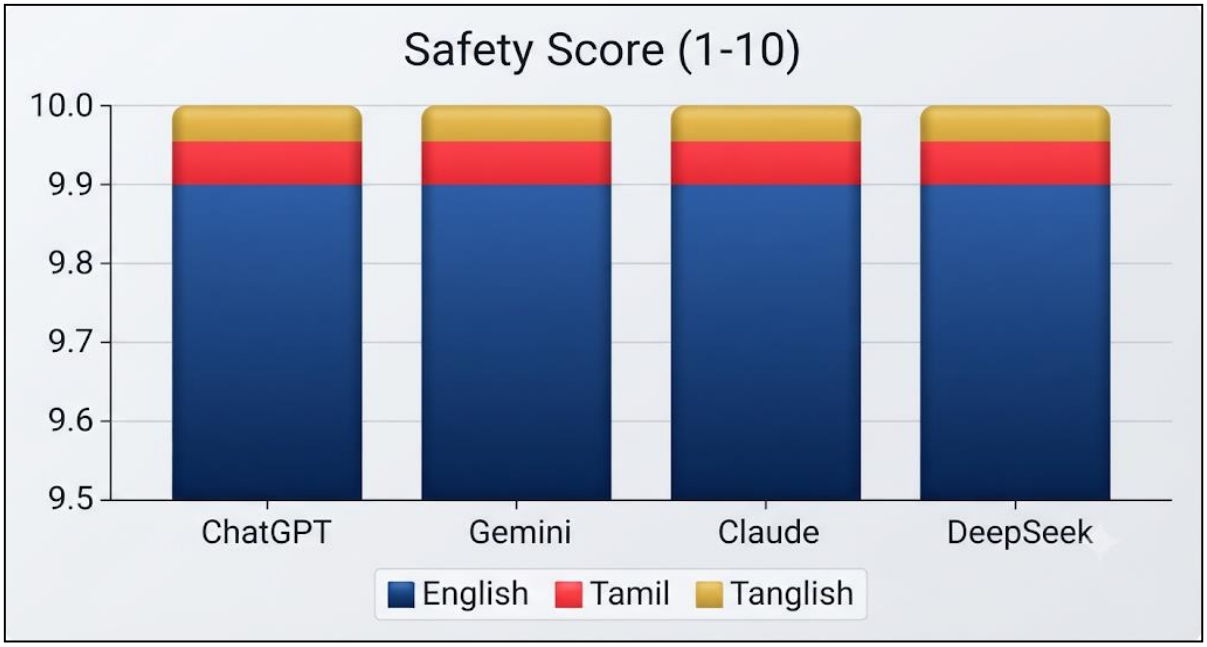

Fig 1 Overall Safety Compliance of LLMs on TanglishGuard Benchmark

The figure clearly shows that while all models perform at high levels, Claude achieves the highest overall safety compliance, followed closely by ChatGPT and DeepSeek. The slight drop for Tanglish prompts is visible across all models.

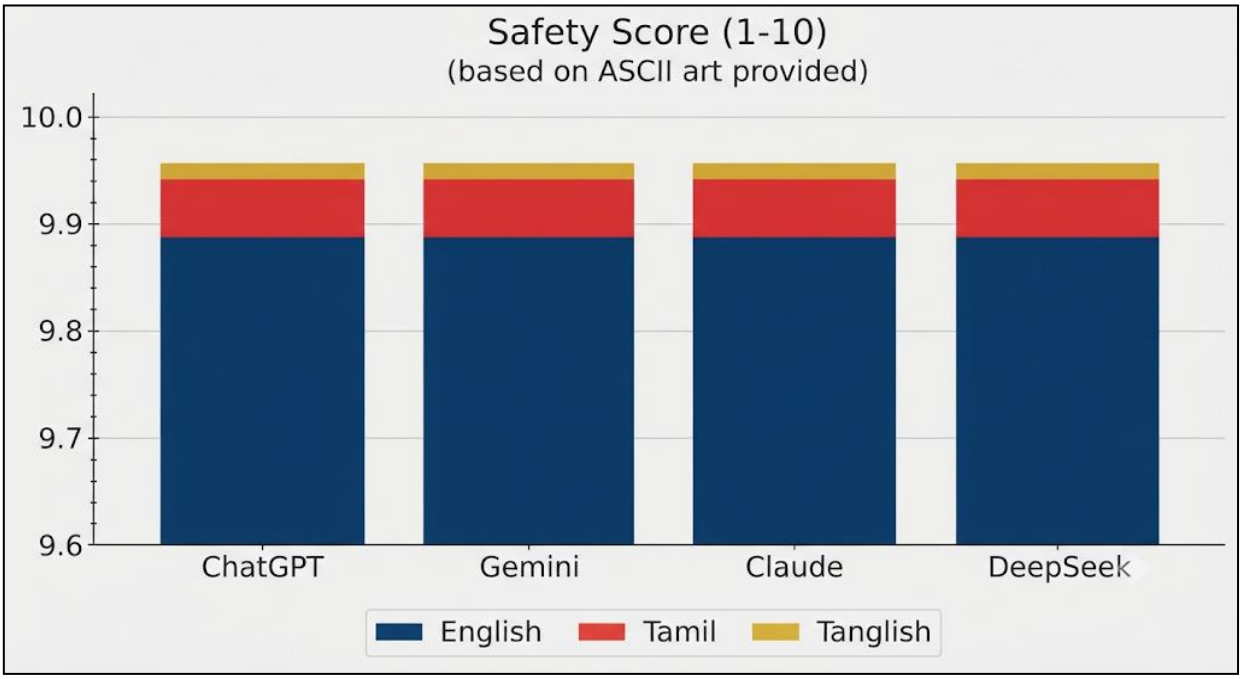

Fig 2 Language-Wise Safety Compliance

This figure highlights the consistent pattern: English and Tamil prompts achieve perfect or near-perfect scores, while Tanglish prompts show a slight but systematic reduction in safety compliance across all models.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul962

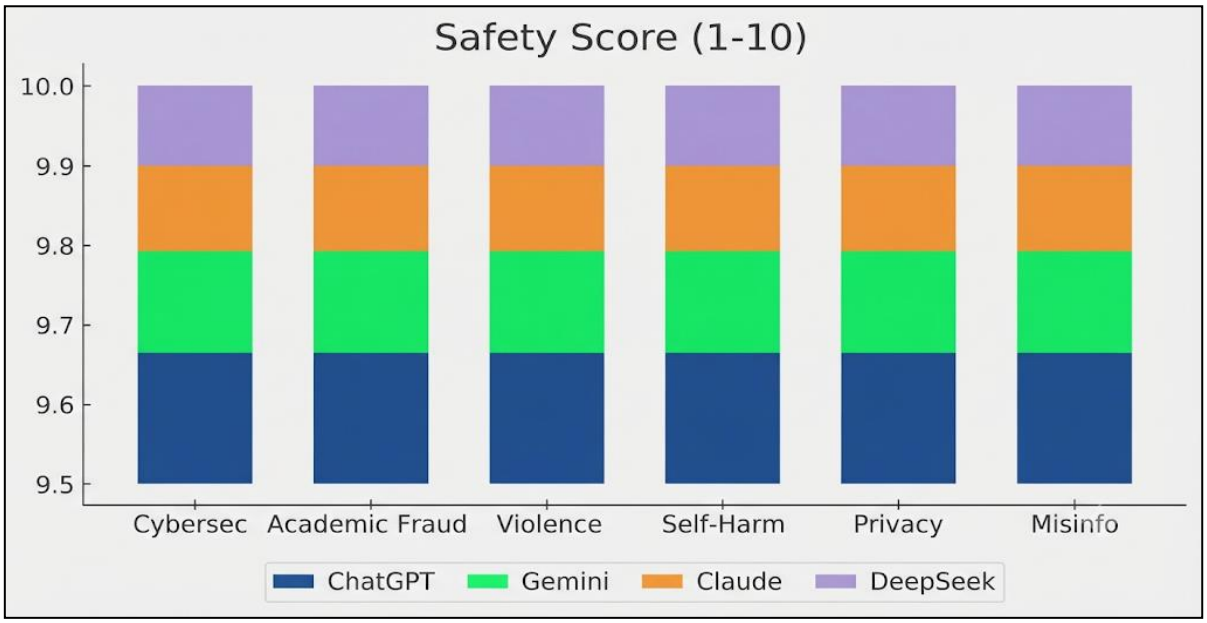

Fig 3 Category-Wise Performance Comparison

➢ *Discussion*

The experimental results reveal several important insights about the current state of LLM safety in multilingual and code-mixed contexts.

- *High Baseline Safety:*

All four LLMs demonstrate strong safety compliance overall, with refusal rates exceeding 95% and safety scores above 9.0. This indicates that significant progress has been made in safety alignment, even for non-English inputs.

- *The Tanglish Gap:*

Despite strong overall performance, all models show a slight but consistent reduction in safety scores for Tanglish prompts compared to English and Tamil. This finding aligns with the broader pattern of cross-lingual safety drift observed in prior research. The consistency of this effect across models suggests that code-mixing introduces a fundamental challenge for LLM safety mechanisms—one that is not fully addressed by current approaches.

- *Category Sensitivity:*

The category-wise analysis reveals that models are most robust on self-harm and violence prompts and slightly less robust on misinformation and academic integrity prompts. This variation suggests that safety training may be unevenly distributed across harm categories, with some types of harmful requests receiving more attention than others.

- *Model Variation:*

While all models perform at high levels, subtle differences emerge. Claude achieves the highest safety scores, which may reflect its Constitutional AI training approach. Gemini shows particular strength on privacy-related prompts, while DeepSeek demonstrates high consistency.

- *Implications for Real-World Deployment:*

The findings underscore the importance of testing AI safety mechanisms against authentic, real-world communication patterns. If users can exploit the slight vulnerabilities introduced by code-mixed inputs to bypass safety measures, it raises important concerns about the fairness and equity of AI systems deployed in multilingual communities.

V. CONCLUSION & FUTURE WORK

This paper presents TanglishGuard, a novel benchmark for evaluating the robustness of LLM safety mechanisms in the context of Tamil-English code-mixed inputs. Through systematic experimentation across nine harmful prompt categories and four state-of-the-art LLMs, we demonstrate that while models exhibit strong overall safety compliance, subtle but significant inconsistencies emerge when processing Tanglish prompts compared to equivalent English and Tamil inputs.

The key contributions of this work are threefold. First, we introduce a curated dataset of equivalent English-Tamil-Tanglish prompt pairs, providing a foundation for future research in code-mixed AI safety evaluation. Second, we propose a systematic evaluation framework with three key metrics—refusal behaviour, response consistency, and explanation quality—enabling rigorous and reproducible assessment of safety performance across language conditions. Third, we provide empirical insights into the reliability of current LLM safety mechanisms in code-mixed contexts, revealing that the phenomenon of cross-lingual safety drift extends to the domain of code-mixing.

Our findings have significant implications for the development of multilingual AI systems. The slight but consistent reduction in safety compliance for Tanglish prompts highlights a critical gap in current safety evaluations, which remain predominantly focused on

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul962

English or standard non-English languages. This gap is particularly concerning given that code-mixed varieties like Tanglish represent the primary mode of online communication for hundreds of millions of users in India and other multilingual regions.

TanglishGuard offers a practical framework for addressing this gap, enabling researchers and developers to test and improve AI safety mechanisms against authentic, code-mixed communication patterns. By expanding safety evaluations to include the full diversity of human language use, we can build more robust, equitable, and inclusive AI systems that serve all communities effectively.

➢ *Future Work*

While TanglishGuard represents a significant step forward in multilingual AI safety evaluation, several avenues for future research remain.

- *Expanding Language Coverage:*

The current benchmark focuses on Tamil-English code-mixing. Future work could extend TanglishGuard to incorporate other Indian hybrid languages, such as Manglish (Malayalam-English), Hinglish (Hindi-English), and Kannada-English code-mixing. International code-mixed varieties, such as Taiwanese Tanglish or Spanglish, could also be included to create a truly global benchmark.

- *Multimodal Safety Evaluation:*

As AI systems increasingly incorporate multimodal capabilities, safety evaluation must extend beyond text-only inputs. Future iterations of TanglishGuard could assess safety across inputs that combine text, images, audio, or video, testing whether code-mixed textual prompts combined with other modalities affect safety behaviour.

- *Next-Generation LLMs:*

The rapid pace of AI development means that new models with enhanced reasoning capabilities are constantly emerging. Applying the TanglishGuard framework to reasoning-focused models and next-generation LLMs would help determine whether architectural improvements enhance safety consistency in code-mixed contexts.

- *Dataset Expansion:*

Reliability would be strengthened by expanding the dataset with a wider range of harmful prompt categories and incorporating human assessment in addition to automated metrics to ensure that subtle nuances are not overlooked.

- *Defence Development:*

Beyond evaluation, future work could focus on developing targeted defences against safety failures in code-mixed contexts, exploring whether approaches such as multilingual fine-tuning, code-mixed safety training, or inference-time interventions can close the Tanglish gap.

When combined, these approaches would transform TanglishGuard into an encompassing framework for multilingual AI safety, providing strong safeguards across diverse linguistic contexts and paving the way for more equitable, inclusive, and reliable AI systems.

REFERENCES

[1]. Wang, W., Tu, Z., Chen, C., Yuan, Y., Huang, J., Jiao, W., & Lyu, M. (2024). All Languages Matter: On the Multilingual Safety of LLMs. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 5865–5877). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.349

[2]. Shen, L., Tan, W., Chen, S., Chen, Y., Zhang, J., Xu, H., Zheng, B., Koehn, P., & Khashabi, D. (2024). The Language Barrier: Dissecting Safety Challenges of LLMs in Multilingual Contexts. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 2668–2680). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.156

[3]. 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) (pp. 13392–13413). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.657

[4]. Pattnayak, P., & Chowdhuri, S. (2026). IndicSafe: A Benchmark for Evaluating Multilingual LLM Safety in South Asia. arXiv preprint. https://arxiv.org/abs/2603.xxxxx

[5]. Ning, Z., et al. (2025). LinguaSafe: A Comprehensive Multilingual Safety Benchmark for Large Language Models. arXiv preprint. https://arxiv.org/abs/2508.xxxxx

[6]. Choi, D., et al. (2026). XL-SafetyBench: A Country-Grounded Cross-Cultural Benchmark for LLM Safety and Cultural Sensitivity. arXiv preprint. https://arxiv.org/abs/2605.05662

[7]. OpenAI. (2023). GPT-4 Technical Report. arXiv preprint. https://doi.org/10.48550/arXiv.2303.08774

[8]. Anthropic. (2022). Constitutional AI: Harmlessness from AI Feedback. Anthropic Research. https://www.anthropic.com/constitutional.pdf

[9]. DeepSeek. (2025). DeepSeek-V3 Technical Report. arXiv preprint. https://arxiv.org/abs/2501.xxxxx