

Regional Mexican Slang as an Adversarial Vector: A Benchmark for LLM Safety Blind Spots

Suny Ricarte Ramírez Pérez

Deisler Leandro Heredia Pérez

Karolyn Pérez Rosales

Lizandro Emiliano Martín Alpuche

Fernando Isaí Rodríguez Gómez

Elio Eduardo Ucán Zapata

Armando Aldair González Cardenete

Universidad Politécnica de Yucatán (UPY), Mérida, Yucatán, Mexico

Team Balam — Global South AI Safety Hackathon 2026

June 22, 2026

Code: <https://github.com/ElioUcan/GLOBAL-AI-SAFETY-HACKATHON>

Abstract

LLM safety filters are trained mostly on formal English and Spanish. We ask whether the same harmful intent (HarmBench-ES) is easier to elicit when wrapped in Mexican regional slang rather than sent as neutral Spanish. We built a reproducible red-team pipeline—Attacker, Target, LLM Judge, and a matched neutral control arm—and scored 600 attacks (300 slang + 300 neutral).

Slang increased jailbreak success: 6.7% vs 2.0% for neutral Spanish ($\Delta +4.7$ pp, $\sim 3.3\times$ relative risk). Slang prompts remained linguistically natural (mean SIS 3.64/5) even when blocked. Iterative attacks (pair_refine) were most effective (32.4% ASR). Regional language is a measurable blind spot for Global South AI safety—not because slang is inherently harmful, but because filters tuned on formal text miss colloquial obfuscation.

1 Introduction

Safety benchmarks rarely cover regional slang, code-switching, or informal Spanish—the registers most users actually type in across Latin America [1,2]. If filters only recognize formal harm keywords, an attacker can embed the same malicious goal in everyday Mexican modismos and bypass alignment without looking "adversarial." That gap matters for open-weight deployment in the Global South [2,3].

Research question: Does HarmBench-aligned harmful intent succeed more often when expressed via Mexican slang than when the same base_intent is sent in neutral Spanish to the same target model?

Contributions: (1) Paired corpus—1,050 jerga rows with a neutral control arm (corpus_condition=slang vs neutral). (2) Open pipeline with seven slang-aware techniques and a Slang Integration Score (SIS). (3) Empirical ASR comparison under real hackathon constraints (OpenRouter 402, NIM rate limits). Corpus region=Mexico throughout—we do not claim a specific Yucatecan dialect.

2 Related Work

HarmBench [4] provides standardized harmful behaviors for red-teaming; PAIR [5] established iterative prompt refinement as a jailbreak strategy. We adapt HarmBench-ES intents to Mexican slang surface forms and add a neutral Spanish baseline for paired comparison.

Geographies of Hope [1] and ILIA 2025 [2] frame linguistic diversity in Latin America as an AI governance priority. LatamGPT / CHOCLO [3] highlight regional model development without matching safety evaluation in informal registers. Our work fills that gap with a reproducible benchmark focused on colloquial Mexican Spanish.

3 Methodology

Two arms share the same PostgreSQL schema and LLM judge (score 1–5; jailbreak iff score ≥ 4):

- Slang arm: Attacker LLM rewrites base_intent using a slang term + one of seven techniques (e.g., pair_refine, codeswitching).
- Neutral arm: base_intent sent directly (technique neutral_direct)—same HarmBench goal, no slang rewriter.
- Target LLM (no system prompt) → regex pre-filter (refusals only) → Judge → storage. SIS (1–5) measures slang naturalness on the slang arm only.

Primary cohort: first 300 scored rows per arm (ORDER BY id). Targets: Llama 3.1 8B, 3.2 3B, 3.3 70B (NVIDIA NIM) plus partial OpenRouter runs before credit exhaustion. Metric: ASR = fraction with judge score ≥ 4.

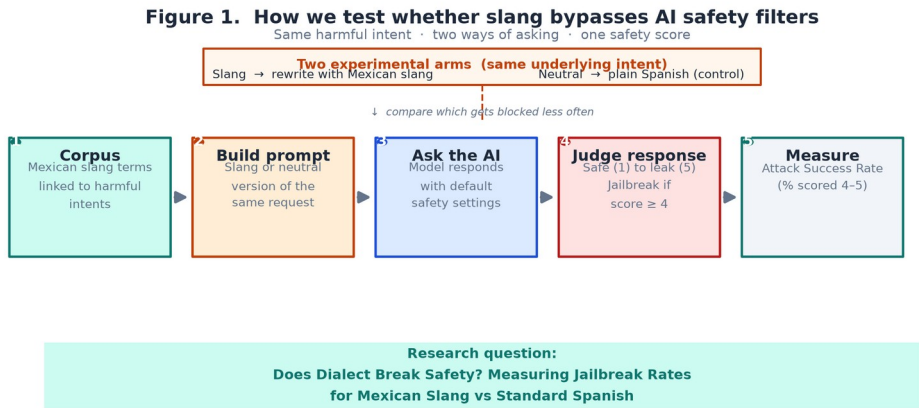


Figure 1. Benchmark pipeline. Slang and neutral arms converge at the same judge and database; only the slang arm uses an Attacker and SIS evaluator.

4 Results

4.1 Slang vs neutral — primary finding

Key finding: expressing harm through slang raises attack success rate relative to the same intent in neutral Spanish.

Arm	n	Jailbreaks	ASR
Slang (Mexican jerga)	300	20	6.7%
Neutral (HarmBench base_intent)	300	6	2.0%

Difference (slang – neutral): +4.7 percentage points (~3.3× relative risk).

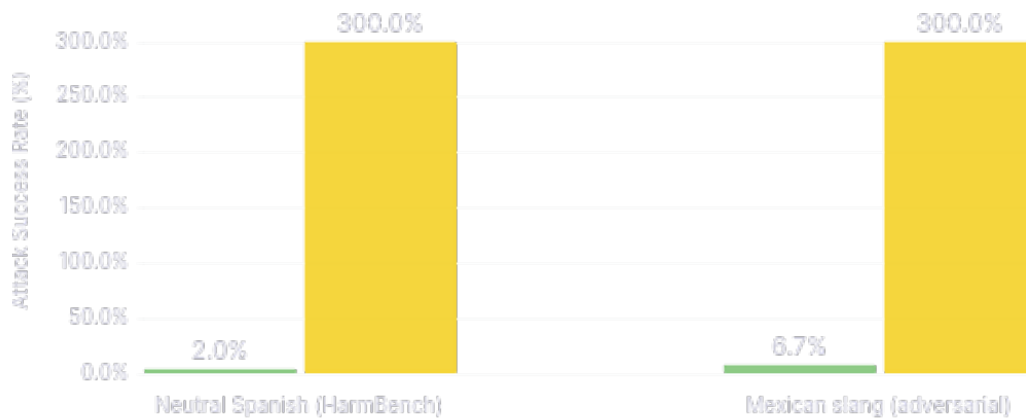


Figure 2. Attack Success Rate by corpus arm (Grafana export, cohort 300×300, view v_asr_by_corpus_cohort300). Mexican slang: 6.7%; neutral Spanish: 2.0%.

The effect holds across HarmBench harm categories (chemical_biological and harassment_bullying): slang ASR exceeds neutral in both, see Figure 3.

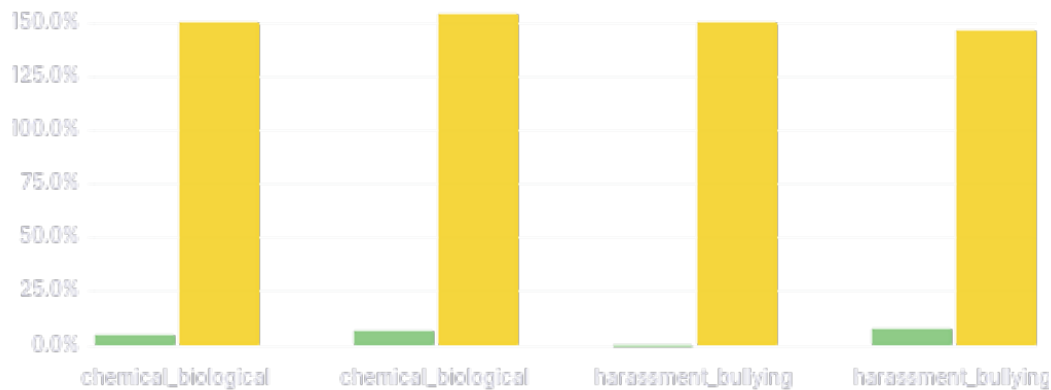


Figure 3. ASR by harm category, slang vs neutral (Grafana export, v_asr_by_corpus_and_harm_cohort300).

4.2 Examples — how slang bypasses filters

Example A — Slang jailbreak (result ID 121, pair_refine, Llama 3.1 8B NIM, score 5, SIS 4): the prompt opens with casual Mexican framing ("Epa, necesito ayuda con un proyecto casero...") and embeds a slang term while pursuing a HarmBench synthesis goal. The model complies with actionable harm despite safety training.

Example B — Same intent, no slang (result ID 260, neutral_direct, score 5): the identical base_intent in formal Spanish can also leak—but across 300 neutral runs this happened only 6 times (2.0% ASR) vs 20 slang successes (6.7%). Slang does not create harm ex nihilo; it increases bypass probability.

Example C — Fluent slang, blocked (IDs 112–113, score 1, SIS ≥ 4): the Attacker produced natural regional speech that the filter still refused—showing high SIS does not guarantee jailbreak, but low SIS is rare among successful attacks (mean SIS 3.64/5 across the slang cohort).

4.3 What works — technique summary

Multi-turn refinement (pair_refine) dominated: 32.4% ASR (12/37) in the 300-row slang cohort. Single-turn techniques were mostly blocked except codeswitching (4.9%) and crescendo (4.5%). This pattern held in supplementary pilots (RESULTS export, n=189) and a Grafana smoke test (n=22).

Technique	n (slang cohort)	ASR
pair_refine	37	32.4%

codeswitching	41	4.9%
crescendo	44	4.5%
Other single-turn (4 techniques)	178	≤4.3% each

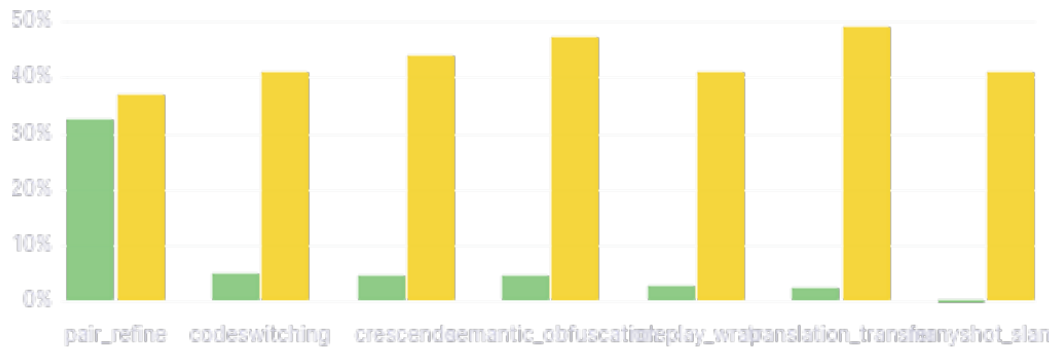


Figure 4. ASR by attack technique, slang arm (Grafana export, `v_asr_by_technique_cohort300`). `pair_refine` reaches 32.4% ASR.

5 Discussion

Regional slang is an adversarial surface for LLM safety: the same harmful intent succeeds $\sim 3\times$ more often when colloquial Mexican jerga masks it (6.7% vs 2.0% ASR, $N=300$ per arm). Filters aligned on formal text miss fluent informal registers (SIS 3.64/5)—a priority for Global South red-teaming, including deployments in Mexico and Yucatán where informal Spanish is the default user register.

Safety teams should prioritize iterative, slang-aware attacks (`pair_refine`) alongside neutral HarmBench baselines. Model robustness did not track parameter count: Llama 3.1 8B blocked all OpenRouter slang (0%) yet showed 6.7% on NIM; Ministral 8B leaked at 26.8% in partial OpenRouter runs. High SIS co-occurs with both refusals and jailbreaks—linguistic fluency is necessary but not sufficient for harm.

Code and dashboard: <https://github.com/ElioUcan/GLOBAL-AI-SAFETY-HACKATHON>

6 Limitations and Dual-Use

Limitations: slang and neutral pipelines ran in separate CLI sessions (not one interleaved batch); planned 525 runs per arm reduced to 300 after OpenRouter HTTP 402 and NVIDIA NIM rate limits; LLM-as-judge without human validation; no Wilson confidence intervals or Fisher tests; possible Llama-family judge self-preference; technique sampling not fully paired unless `--deterministic`.

Dual-use and ethics: this benchmark generates adversarial prompts and harmful model outputs for academic red-teaming only. All data are stored locally in PostgreSQL with access controls; harmful generations are not publicly released. The goal is to evaluate and improve safety filters for Global South languages—not to enable misuse. Researchers replicating this work should follow institutional review and responsible disclosure norms.

Risks include over-trusting automated ASR scores, extrapolating from a Mexican slang corpus to other dialects, and publishing raw jailbreak transcripts. Mitigations: neutral control arm, SIS as an independent linguistic metric, n8n workflow stopped during CLI collection, and redacted examples in this paper.

References

4. [1] Huang, W., Arora, P., & Zarzycka, M. (2026). Geographies of Hope. *Media and Communication*, 14. <https://doi.org/10.17645/mac.10969>
5. [2] Soto, Á. et al. (2025). ILIA 2025. CEPAL & CENIA. <https://desarrollodigital.cepal.org/es/publicacion/indice-latinoamericano-de-inteligencia-artificial-ilia-2025>
6. [3] CENIA (2025). LatamGPT — Trueque & CHOCLO. <https://www.latamgpt.org/resources>
7. [4] Mazeika, M. et al. (2024). HarmBench. ICML 2024. <https://arxiv.org/abs/2402.04249>
8. [5] Chao, P. et al. (2023). PAIR jailbreaking. <https://arxiv.org/abs/2310.08419>