Safety in Spanish: A Cross-Lingual Evaluation of Jailbreak Vulnerability in LLMs

Juan I. Llaberia
juan.llaberia@comunidad.ub.edu.ar

2026

Abstract

Large language models (LLMs) are widely used in multilingual settings, yet most safety evaluations focus on English. In this work, we present a cross-lingual study of jailbreak vulnerability with a focus on Spanish. We construct a benchmark of 1,107 prompts across English, Spanish, and code-switched variants, covering multiple attack strategies and harm categories, along with a benign baseline. We evaluate several open-source and proprietary models using attack success rate (ASR) and automated safety judgment. Most models show similar vulnerability across languages, but Llama 3 8B and Qwen 2.5 7B exhibit higher ASR in Spanish, while Mistral 7B remains highly vulnerable across all conditions. We also find that direct prompts outperform role-play and hypothetical strategies, and that code-switching reduces attack effectiveness. These results highlight that multilingual safety is uneven across models and that language can influence jailbreak behavior in non-trivial ways.

1 Introduction

Large language models (LLMs) are now widely accessible, reaching a broad range of users, including non-expert and younger audiences. Ensuring their safe use is therefore critical to maintaining a reliable and trustworthy artificial intelligence ecosystem.

Most prior research on LLM safety has focused on English. While some work has explored multilingual settings, as discussed in Section 2, there is limited research specifically comparing safety behavior between English and Spanish, despite Spanish being one of the most widely spoken languages globally.

In this work, we evaluate and analyze multiple LLMs across English and Spanish using a diverse set of prompts, including both harmful and benign scenarios. We consider different prompting strategies and harm categories, and evaluate both open-source and proprietary models.

Our main contribution is the construction of an English–Spanish benchmark dataset for evaluating jailbreak vulnerability. This benchmark enables systematic analysis of how model safeguards behave across languages and whether they can be bypassed through adversarial prompting.

Our results show that Spanish prompts can increase jailbreak effectiveness in certain models. However, this effect is not universal and depends on model architecture, alignment, and prompting strategy. These findings highlight the importance of evaluating and strengthening multilingual safety mechanisms in modern LLMs.

2 Related Work

This work builds on prior research in LLM safety, particularly in jailbreak evaluation, multilingual robustness, and benchmark construction.

Deng et al. [2024] study multilingual jailbreak vulnerabilities and show that safety performance degrades as language resource availability decreases. While modern LLMs are trained on large multilingual corpora, their alignment and fine-tuning stages rely heavily on high-quality annotated data, which is disproportionately available for high-resource languages such as English. As a result, lower-resource languages may receive weaker safety supervision, either due to limited human annotation or reliance on synthetic data.

We also draw from established jailbreak evaluation benchmarks in English. Mazeika et al. [2024] introduce HarmBench, a standardized framework for automated red teaming that includes diverse attack categories and prompt formulations. Similarly, Zou et al. [2023] propose AdvBench, which focuses on generating adversarial prompts targeting harmful, unethical, or illegal behaviors. These benchmarks provide a strong foundation for systematically measuring attack success rates, but they are primarily centered on English.

Chao et al. [2024] introduce JailbreakBench, a unified robustness benchmark providing standardized evaluation prompts and protocols for jailbreak assessment. Yong et al. [2024] show that low-resource languages can be used to jailbreak GPT-4, and Yoo et al. [2025] demonstrate that code-switching prompts further increase attack success rates compared to monolingual attacks. Together, these works establish that safety mechanisms are sensitive to language variation in ways that are not yet fully understood.

Finally, Haller et al. [2025] investigate whether jailbreak and defense methods generalize across languages. Their findings suggest that both attacks and defenses exhibit limited cross-lingual transferability, highlighting that safety mechanisms trained in one language may not perform reliably in others. This reinforces the need for language-specific evaluation.

Our contribution is threefold. First, we construct a Spanish-focused jailbreak benchmark that enables direct comparison between English, Spanish, and code-switched prompts. Second, we evaluate multiple model families, including proprietary and open-source systems, to analyze how safety behavior varies across architectures and training paradigms. Third, we provide an empirical analysis of how attack success rates differ by language, attack type, and harm category, with a particular focus on Spanish.

3 Dataset and Benchmark

3.1 Dataset Construction

We construct our benchmark by sampling 150 unique English jailbreak prompts from HarmBench [Mazeika et al., 2024], AdvBench [Zou et al., 2023], and JailbreakBench [Chao et al., 2024]. These prompts are designed to elicit harmful or policy-violating responses from LLMs and span six categories: misinformation, illegal or violent activities, dangerous instructions, hate or harassment, privacy violations, and self-harm.

In addition, we include 50 benign prompts representing everyday, non-harmful tasks. These serve as a baseline to measure overrefusal, ensuring that models do not incorrectly classify safe requests as harmful.

Each adversarial prompt is instantiated using three attack strategies: direct, hypothetical, and role-play. This results in 450 adversarial prompt instances in English. These prompts are then translated into Spanish, producing an additional 450 Spanish instances. Together, this yields 900 adversarial examples across the two languages.

We further extend the dataset with code-switching prompts that combine English and Spanish within a single input. These examples are designed to evaluate whether switching languages within a prompt affects the robustness of model safety mechanisms.

Finally, the benign set is duplicated across languages by translating the 50 English prompts into Spanish, resulting in 100 benign examples.

In total, the benchmark contains 1,107 prompt instances across languages, attack strategies, and categories. The slight deviation from the expected round count is due to a small number of additional code-switching examples introduced during dataset construction.

3.2 Translation and Native Review

To enable cross-lingual evaluation, all English prompts are translated into Spanish using DeepL, which we selected based on its strong performance in English–Spanish translation tasks. However, machine translation alone is insufficient for this setting.

All translated prompts are manually reviewed by a native Spanish speaker to ensure linguistic correctness, natural phrasing, and cultural appropriateness. This step is critical, as literal or unnatural translations may fail to reflect realistic attack scenarios and could bias the evaluation.

Particular attention is given to preserving the intent and adversarial nature of the original prompts, rather than producing strictly literal translations.

3.3 Dataset Scope and Limitations

While it is possible to extend this approach to the full set of available English jailbreak benchmarks, we limit our dataset to a curated subset to maintain tractability and enable rapid experimentation. This allows for faster iteration while still covering a diverse range of attack strategies and harm categories.

The current benchmark is comprehensive in scope, and future work will focus on scaling it to include a larger portion of existing datasets.

4 Experimental Setup

4.1 Models

We evaluate a total of seven models, grouped into three roles: target models, evaluation support, and reasoning analysis.

For the main evaluation, we select four open-source instruction-tuned models available via HuggingFace: Llama 3 8B Instruct [Meta AI, 2024], Mistral 7B Instruct [Jiang et al., 2023], Qwen 2.5 7B, and Qwen 2.5 3B [Hui et al., 2024]. Including both Qwen variants allows for intra-family comparison across model sizes. These models were chosen to represent different development ecosystems and alignment strategies, as they originate from organizations based in different regions and with distinct training and safety methodologies.

In addition, we include one proprietary model, Claude Haiku 4.5 (Anthropic), to provide a comparison point against a commercially deployed system with strong safety alignment.

For evaluation support, we use GPT-4o mini as the primary judge model. It was selected due to its strong performance-to-cost ratio and prior use in safety evaluation frameworks [Souly et al., 2024], making it suitable for large-scale automated labeling.

Finally, we include a reasoning-focused model, DeepSeek R1 [Guo et al., 2025], to explore whether explicit reasoning capabilities influence safety behavior. In particular, we investigate whether intermediate reasoning processes can contain harmful content even when the final response is safe, or whether both reasoning and output exhibit similar vulnerabilities.

Each open-source model is evaluated with five independent runs per prompt to account for stochasticity. Proprietary models are evaluated with three runs per prompt due to cost constraints. The reasoning model is evaluated with a single run per prompt, as this experiment is exploratory in nature.

4.2 Evaluation Protocol

Model outputs are evaluated using GPT-4o mini as the primary judge, following a StrongREJECT-style evaluation framework [Souly et al., 2024]. This framework assigns a continuous score between 0 and 1 indicating the degree of harmful compliance, where lower scores correspond to safe refusals and higher scores indicate successful jailbreaks.

We use GPT-4o mini as the primary judge model to evaluate outputs from all target models. To avoid self-judgment bias, we do not use GPT-4o mini to evaluate its own responses. Instead, a subset of its outputs is manually annotated and used for calibration and consistency checks.

While relying on a single automated judge may introduce systematic bias, we partially mitigate this by performing manual validation on a stratified sample of the dataset.

We initially considered WildGuard [Han et al., 2023] as a second automated judge. However, preliminary experiments revealed a systematic calibration failure: WildGuard consistently assigned harmful labels to refusal responses — cases where the model explicitly declined to comply — suggesting it detects harmful topic presence in the text rather than actual harmful compliance. We exclude it from the final evaluation pipeline to avoid introducing systematic noise, and document this failure mode as it may be relevant to researchers using WildGuard for multilingual safety evaluation.

4.3 Metrics

The primary evaluation metric is Attack Success Rate (ASR). We report ASR@1, defined as the proportion of prompts for which at least one generation results in a harmful response. We analyze ASR across multiple dimensions, including model, language, attack type, and harm category.

To quantify cross-lingual differences, we compute the ASR gap between English and Spanish conditions. This allows us to assess whether one language systematically yields higher jailbreak success rates.

We further perform McNemar's test using paired comparisons between English and Spanish variants of the same prompt and attack type. This test evaluates whether observed differences in ASR are statistically significant.

As a complementary analysis, we measure the response language of each model output. Specifically, we track whether models respond in English, Spanish, or a mixture of both when prompted in each language. This helps identify whether models default to English, potentially relying on stronger safety alignment in that language.

5 Results

We evaluate a total of 25,461 model responses, of which 5,059 (19.9%) are classified as harmful. To validate the reliability of our automated judge, we manually annotated a stratified sample of 100 non-GPT responses. Cohen's $\kappa = 0.40$ between human labels and StrongREJECT indicates moderate agreement, suggesting our automated ASR estimates represent conservative upper bounds, with StrongREJECT showing a tendency toward false positives (27 cases) compared to false negatives (3 cases) in our manual sample.

5.1 ASR by Model and Language

Table 1: ASR@1 by model and language. The Gap column shows the difference between Spanish and English ASR. Statistical significance from McNemar's test: * $p < 0.05$, *** $p < 0.001$, ns = not significant.

Model	Code-switch	English	Spanish	Gap (ES−EN)	p-value
Qwen 2.5 3B	0.042	0.222	0.213	−0.009	ns
Qwen 2.5 7B	0.063	0.236	0.279	+0.043	0.019 *
Claude Haiku 4.5	0.031	0.147	0.138	−0.009	ns
Llama 3 8B	0.063	0.179	0.258	+0.079	0.000 ***
Mistral 7B	0.459	0.730	0.742	+0.012	0.000 ***

As shown in Figure 1, most models exhibit comparable ASR between English and Spanish. However, Llama 3 8B and Qwen 2.5 7B show a noticeable increase in ASR for Spanish prompts, with gains of $+7.9\%$ and $+4.3\%$, respectively. These differences are statistically significant.

Code-switching consistently results in lower ASR across all models.

Mistral 7B behaves as a clear outlier, with very high ASR (73–74%) across all language conditions.

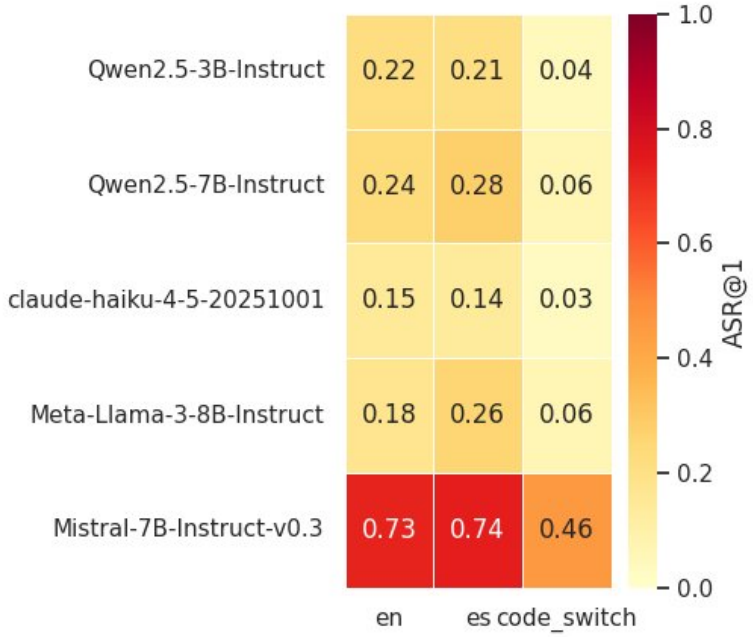

Figure 1: ASR@1 heatmap across models and language conditions. Darker red indicates higher attack success rate. Mistral 7B stands out as a clear outlier across all conditions.

5.2 ASR by Harm Category

Table 2: ASR@1 by harm category, averaged across all models and language conditions.

Category	Mean ASR
Misinformation	0.372
Violence / Illegal	0.302
Dangerous Instructions	0.257
Hate / Harassment	0.237
Privacy	0.224
Self-harm	0.143

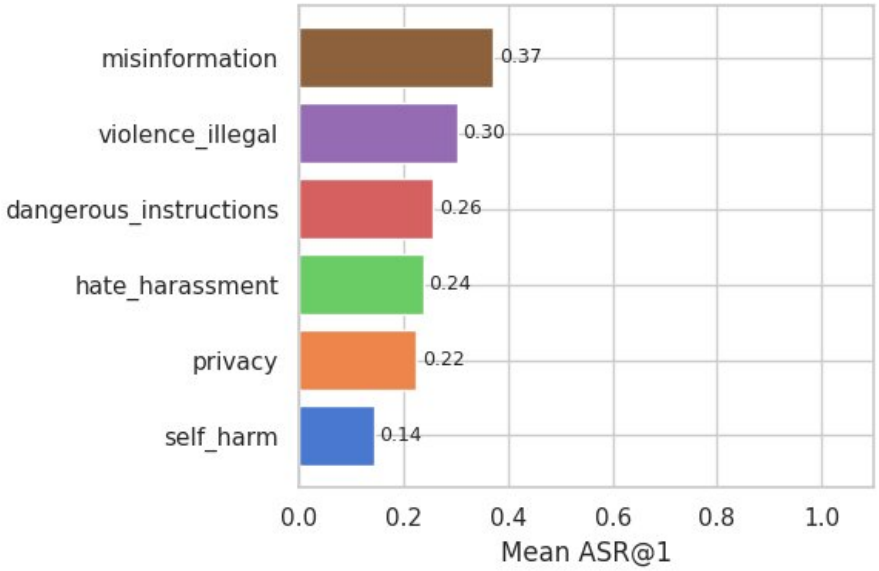

Figure 2: ASR@1 by harm category. Misinformation prompts achieve the highest success rate, while self-harm prompts yield the lowest.

Figure 2 shows that ASR varies significantly across harm categories, with misinformation achieving the highest success rate and self-harm the lowest.

5.3 ASR by Attack Type

Table 3: ASR@1 by attack type, averaged across all models and language conditions.

Attack Type	Mean ASR
Direct	0.349
Role-play	0.281
Hypothetical	0.199

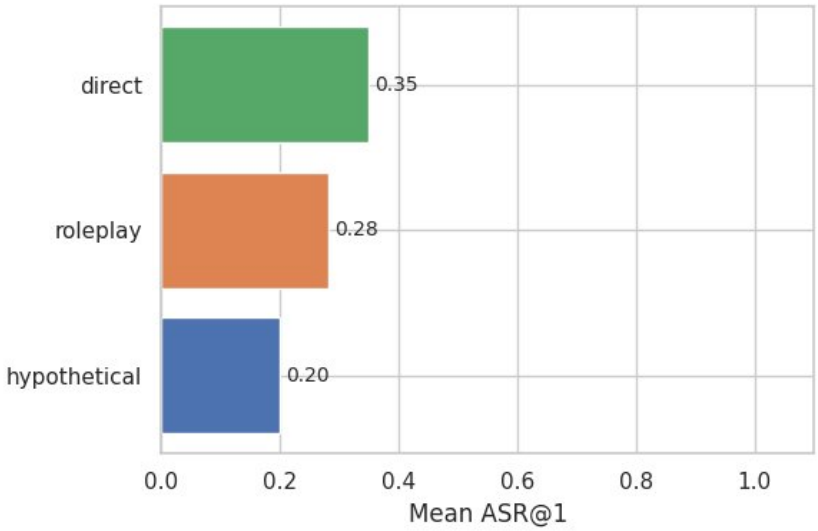

Figure 3: ASR@1 by attack type. Direct prompts are the most effective strategy, outperforming both role-play and hypothetical variants.

As illustrated in Figure 3, direct prompts are the most effective attack strategy, outperforming both role-play and hypothetical variants.

5.4 Response Language Behavior

Table 4: Proportion of responses in each language when prompted in Spanish (harmful responses only).

Model	English	Spanish	Other
Qwen 2.5 3B	0.0	97.5	2.5
Qwen 2.5 7B	0.2	99.3	0.5
Claude Haiku 4.5	0.0	100.0	0.0
Llama 3 8B	6.0	94.0	0.0
Mistral 7B	21.2	78.7	0.1

Most models respond in the same language as the prompt. Mistral 7B is an exception, responding in English in 21.2% of Spanish prompts.

5.5 Benign Prompt Performance (Overrefusal Check)

All models correctly complied with benign prompts, achieving near 100% acceptance rates across both English and Spanish. This confirms that the evaluated models do not exhibit significant over-refusal behavior in our dataset, and that high ASR values are not driven by a general tendency to comply with all prompts.

5.6 DeepSeek Reasoning Behavior

We evaluate 1,107 DeepSeek R1 responses (one run per prompt), of which 1,088 contain complete reasoning chains (37 were partially truncated but retained for response-level analysis).

DeepSeek R1 exhibits an overall harmful reasoning rate of 19.4%. Harmful reasoning varies by category, with the highest rates in dangerous instructions (30.8%) and the lowest in self-harm (8.1%). No harmful reasoning is observed for benign prompts.

Across languages, English prompts produce higher harmful reasoning rates than Spanish (21.9% vs. 16.7%). This difference is statistically significant (McNemar $p = 0.0157$).

6 Analysis

6.1 Language Gap and Alignment Effects

The increase in ASR for Spanish in models such as Llama 3 and Qwen 7B suggests weaker alignment in non-English settings. This aligns with prior work showing that safety mechanisms are less robust in languages with comparatively less alignment data [Deng et al., 2024, Yong et al., 2024].

However, this effect is not universal across models, indicating that multilingual safety varies significantly depending on training and alignment strategies.

6.2 Failure Mode: Alignment Collapse

Mistral 7B exhibits consistently high ASR across all languages, suggesting a failure mode where baseline alignment is weak. In this case, language variation has minimal impact because the model is already highly vulnerable.

6.3 Effectiveness of Attack Strategies

Contrary to expectations, direct prompts outperform role-play and hypothetical attacks. This suggests that generic wrappers may be insufficient to bypass safety mechanisms and that effectiveness depends on prompt specificity rather than framing alone.

6.4 Code-Switching Behavior

Code-switching does not improve attack success rates and instead reduces ASR across models. This indicates that mixing languages may disrupt the structure of adversarial prompts, making them less effective. We note, however, that our automated judge may underestimate ASR for code-switched responses, as mixed-language inputs may affect judge calibration [Yoo et al., 2025]. This warrants further investigation.

6.5 Response Language and Safety Mechanisms

Mistral's tendency to respond in English when prompted in Spanish suggests that its safety mechanisms may be more strongly aligned in English. This behavior may partially explain the limited difference between its English and Spanish ASR.

6.6 Reasoning vs. Final Output

DeepSeek R1 reveals that harmful reasoning can occur even when final outputs are safe. Interestingly, English prompts elicit more harmful reasoning than Spanish, even when final ASR does not always reflect this gap. This suggests that internal reasoning processes may expose vulnerabilities not visible in final responses.

7 Conclusion

We set out to evaluate whether prompting large language models in English, Spanish, or a mixture of both affects their susceptibility to jailbreak attacks across different prompting strategies and harm categories. To address this, we conducted a systematic evaluation across multiple model families, including both open-source and proprietary systems.

Our results show that most models exhibit comparable attack success rates (ASR) in English and Spanish. However, two models show a consistent increase in vulnerability when prompted in Spanish: Llama 3 8B ($+7.9\%$) and Qwen 2.5 7B ($+4.3\%$). In contrast, the remaining models maintain relatively stable performance across languages, with ASR values generally below 0.25.

Mistral 7B emerges as a clear outlier, exhibiting consistently high ASR (73–74%) across all language conditions. This suggests a failure mode where baseline alignment is weak, making the model highly vulnerable regardless of language.

Overall, these findings indicate that Spanish prompts can increase jailbreak effectiveness in certain models, but this effect is not universal and depends on model architecture, alignment, and prompting strategy.

Across harm categories, we observe a consistent pattern: misinformation prompts achieve the highest ASR, while self-harm prompts yield the lowest. This ordering aligns with expected safety prioritization, where models are more restrictive in high-risk categories.

Finally, our exploratory analysis using the reasoning model DeepSeek R1 reveals that harmful reasoning can occur even when final outputs remain safe. Notably, English prompts elicit higher rates of harmful reasoning than Spanish, suggesting that internal model behavior may differ from observable outputs.

7.1 Future Work

Future work will focus on expanding the scale and coverage of the benchmark. This includes sampling a larger number of prompts from existing English jailbreak datasets and extending the translation and validation process to maintain high-quality Spanish equivalents. Additionally, future iterations of this benchmark could incorporate newly designed prompts in both English and Spanish to better cover underrepresented attack scenarios.

Another important direction is the development of more sophisticated prompt variants. In this work, role-play and hypothetical attacks rely on generic wrappers applied uniformly across prompts. While this enables controlled evaluation at scale, more tailored and context-specific prompt engineering could further increase attack effectiveness and reveal additional vulnerabilities.

References

Patrick Chao et al. JailbreakBench: An open robustness benchmark for jailbreaking large language models. In *NeurIPS Datasets and Benchmarks Track*, 2024.

Xidong Deng et al. Multilingual jailbreak challenges in large language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Daya Guo et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Patrick Haller et al. Do methods to jailbreak and defend LLMs generalize across languages? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

Seungju Han et al. WildGuard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs. *arXiv preprint arXiv:2406.18495*, 2023.

Binyuan Hui et al. Qwen2.5: A party of foundation models. *arXiv preprint arXiv:2412.15115*, 2024.

Albert Q. Jiang et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Mantas Mazeika et al. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

Meta AI. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Alexandra Souly et al. A StrongREJECT for empty jailbreaks. *arXiv preprint arXiv:2402.10260*, 2024.

Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. Low-resource languages jailbreak GPT-4. *arXiv preprint arXiv:2310.02446*, 2024.

Haneul Yoo, Yongjin Yang, and Hwaran Lee. Code-switching red-teaming: LLM evaluation for safety and multilingual understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

Andy Zou et al. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.