S-OWMI: A Perimeter Auditing Framework for Open-Weight Models in Latin American Institutions

Ramiro Carnicer Souble
Independent Researcher

Mauricio Genta
Independent Researcher

Federico Hörl
Universidad Nacional de San Martín (UNSAM)

Nicolás Adrian Oroz
Tech del Fuego

With Apart Research

github.com/fede-h/SOWMI

Abstract

Is a model that is "safe" in English equally safe in Spanish? For institutions in the Global South adopting open-weight LLMs, the answer is consequential—and, until now, largely unmeasured. Organizations in the region are increasingly drawn to open-weight large language models (LLMs) to ensure data sovereignty, adapt to local regulatory requirements, and eliminate recurring foreign-currency subscription fees. Yet local hosting and fine-tuning carry substantial infrastructure costs, making pre-deployment auditing essential: institutions cannot afford to commit scarce GPU resources to models with unverified safety profiles in their target language. We introduce the **Spanish Open-Weight Maturity Index (S-OWMI)**, an auditable perimeter evaluation framework for open-weight LLMs. This initial release deliberately scopes to *Vertical 1: Spanish Upstream Data Curation*, providing a clean, non-overlapping safety signal ahead of fine-tuning; integration of *Vertical 2 (Provenance)* and *Vertical 3 (Tiered Release)* is reserved for future work. Evaluating Llama-3.1-8B and Qwen2.5-7B on an English baseline versus an "español-diverso" dataset—mixing neutral Spanish, Spanglish, and regional colloquialisms— we find that dangerous-prompt refusal (ΔASR) and over-refusal (FRR) transfer robustly across languages. The critical finding lies elsewhere: both models show a **+13.3 pp degradation in bias-consistency** when switching to Spanish (step 5, medium Cohen's h), and both score poorly on open-domain LATAM factual queries ($\sim$80% of answers fall short of detailed answer keys, though only 7% (7/100) are outright factual errors—the rest being incomplete) regardless of language—exposing a general readiness gap that a Spanish-only audit would miss without an English baseline. The L1/L2/L3 scorecard gives Global South institutions an actionable, evidence-based tool to surface these weaknesses before committing to local deployment.

1. Introduction

When an institution in Latin America—such as a hospital, fintech, or government agency—seeks to adopt AI, the stakes are concrete and immediate. The regional AI market reached $29.5 billion in 2025 and is expanding at over 26% annually, with banking, healthcare, and government among the fastest-growing sectors. Yet the dominant adoption path—proprietary API subscriptions—creates structural dependency: costs are denominated in foreign currency, data must cross borders, and compliance with national data protection frameworks becomes legally precarious. Brazil's LGPD, Argentina's Ley 25.326, and Colombia's Ley 1581 all impose strict cross-border data transfer requirements; for organizations handling sensitive patient, financial, or citizen data, routing inferences through foreign infrastructure is not merely expensive—it may be legally untenable. Open-weight models resolve this: they allow full local data hosting, enabling regulatory compliance while eliminating recurring foreign-currency fees. However, local hosting and fine-tuning require significant GPU capital expenditure, which is scarce and

expensive in the region. This infrastructure constraint makes pre-deployment perimeter auditing vital—organizations cannot afford to commit scarce compute resources to models with unverified safety profiles in their target language.

Currently, many deployers assume that safety benchmarks conducted in English guarantee secure operation in Spanish. Mechanistically, cross-lingual safety alignment is mediated by a sparse translingual parameter subset—less than 0.3% of global model parameters— known as Shared Safety Neurons (SS-Neurons) [6]. Because Spanish is a high-resource language, it benefits from partial safety transfer through these neurons; dangerous-prompt refusal does not simply collapse when the language switches. However, the SS-Neuron pathway is calibrated primarily on standard, written English and Peninsular Spanish [5]. Regional colloquialisms, code-switching (Spanglish), and Latin American dialectal variation sit below its reliable activation threshold, creating a subtler but operationally significant gap: not outright refusal bypass, but degraded consistency in bias detection and factual reliability—as corroborated by multilingual safety benchmarks [1, 3] and precisely the failure modes our evaluation surfaces. Furthermore, open-weight models present a compounding risk: their built-in safety alignment is fragile and can be surgically removed (abliteration) or inadvertently eroded during local fine-tuning, making pre-deployment auditing in the target language variety essential. Where the SS-Neuron account is *mechanistic*, S-OWMI is its *behavioral* counterpart: it empirically measures, at the deployment boundary, exactly the dialect-level failures that account predicts—and our results refine it, showing that refusal transfers robustly while bias-consistency and factual reliability are the dimensions that actually degrade.

Our core theory of change is that institutions need an auditable and evidence-based way to evaluate Spanish data curation safety *before* adopting and fine-tuning an open-weight model with sensitive local data. By restricting our framework to Vertical 1, we eliminate scoring overlaps with downstream testing and provide a pure perimeter audit.

The S-OWMI Three-Vertical Framework. S-OWMI structures evaluation across three complementary *verticals* (V1–V3); within each vertical, evaluation is organized into numbered *steps*. (We use "vertical" for the three top-level axes and "step" for their components throughout.) **Vertical 1 — Spanish Upstream Data Curation** (implemented here) audits whether a model was trained, filtered, and curated treating Spanish as a first-class safety dimension—covering token fertility, dialectal coverage, over-refusal, jailbreak probes, bias consistency, and factual quality in Spanish. **Vertical 2 — Weight Provenance & Robustness** (future work) addresses model traceability and post-adaptation safety: identity verification (hashes, versioning), weight watermarking, and survival of safety constraints after fine-tuning, LoRA, quantization, and abliteration. **Vertical 3 — Spanish Tiered Release & Testing** (future work) verifies that the model underwent adversarial red-teaming in Spanish prior to deployment, covering jailbreaks, prompt injection, RAG poisoning, code-switching evasion, dialectal stress-testing, and native-speaker red-teaming. This paper implements Vertical 1 in full; Verticals 2 and 3 are reserved for future work to avoid scoring overlaps.

Our main contributions are:

1. **S-OWMI:** An auditable safety index focusing on Spanish data curation for open-weight models.
2. **Empirical Evaluation:** A comparative analysis of Llama-3.1-8B and Qwen2.5-7B, revealing robust refusal-safety transfer alongside a consistent bias-consistency degradation in dialectal Spanish and a shared factual accuracy gap in LATAM-specific domains.
3. **Organizational Scorecard:** A practical L1/L2/L3 rubric that helps Global South organizations interpret empirical results to make informed adoption decisions.

2. Related Work

Our framework builds upon recent multilingual safety literature while addressing critical gaps for Latin America:

- **M-ALERT [1] (2024):** Demonstrated high safety inconsistency across European languages, showing that harmful prompts rejected in English are often fulfilled in standard Spanish.
- **LinguaSafe [2]:** A benchmark evaluating safety across 12 languages. However, it lacks Latin American Spanish, dialectal variations, and an organizational maturity scorecard.
- **PolygloToxicityPrompts [3] (PTP):** Demonstrated that toxicity increases in languages with fewer high-quality data resources, showing that safety alignment is strongly correlated with the volume and curation of training data in each language. This directly justifies the necessity of S-OWMI's focus on upstream curation.
- **MultiJail [4] (Marx & Dunaiski, 2026):** Using low-resource African languages (Afrikaans, Kiswahili, isiXhosa, isiZulu), demonstrated jailbreak success rates of 59.8% via automated translation and 75.8% via human red-teaming—confirming that cross-lingual safety gaps widen as language resource level decreases.
- **"Spanish Is Not Just One" [5]:** Proved empirically that LLMs have an implicit bias toward Peninsular Spanish and treat regional variants differently.
- **Sycophancy as a Multilingual Alignment Failure [7] (Shah et al., 2026):** Showed that alignment behavior degrades across languages, topics, and models—reinforcing that cross-lingual robustness cannot be assumed and motivating the exact premise S-OWMI operationalizes for Latin American Spanish.

Unlike existing benchmarks, S-OWMI combines Latin American variations ("español-diverso") with a practical governance and scoring framework for open-weight models.

3. Methods

The S-OWMI framework evaluates whether an open-weight model has been trained, filtered, and curated treating Spanish as a first-class safety dimension, or if it merely acquired Spanish capabilities passively. We focus on **Vertical 1: Spanish Upstream Data Curation**, which establishes an auditable perimeter defense before deployment.

3.1 The Full 6-Step S-OWMI Vertical 1 Protocol

A comprehensive upstream data curation audit covers six distinct linguistic and safety dimensions:

1. **Linguistic Distribution of the Base Corpus (Step 1):** Token Fertility Analysis — the tokenizer processes parallel corpora in English and Spanish to calculate the compression ratio (characters per token). High fertility indicates poor training data representation. *Metric: Spanish Corpus Coverage (SCC).*
2. **Dialectal Coverage and Blind Spots (Step 2):** Harmful prompts in neutral Spanish are perturbed across 21 Spanish-speaking nations and common code-switching (Spanglish). *Metric: Regional Coverage Score (RCS).*
3. **Semantic Filtering and Over-refusal (Step 3):** Model is prompted with benign queries containing high-sensitivity regional triggers. High over-blocking signals reliance on translated keyword lists. *Metric: False Refusal Rate (FRR).*

4. **Dangerous Knowledge Probes (Step 4):** Prompts injected under three adversarial vectors: semantic evasion, code-switching, and visual formatting. *Metric: Attack Success Rate Delta (ΔASR).*
5. **Toxicity, Biases, and Local Stereotypes (Step 5):** Responses to ambiguous, locally-framed prompts are judged on whether they *challenge*, stay *neutral* toward, or *reinforce* a biased premise. *Metric: Bias and Stereotype Score (BSS) — the fraction of responses that challenge rather than reinforce the premise.*
6. **Factual Quality and Local Hallucinations (Step 6):** Benchmarked using localized medical, financial, and legal tasks. *Metric: Factuality Evaluation Score (FES) and False Citation Rate (FCR).*

3.2 MVP Scope and Implementation

To build an immediate, executable auditing tool for Latin American organizations, our MVP focuses on Steps 3–6, which have the highest signal-to-noise ratio and organizational risk relevance. Step 1 (token fertility) is additionally computed and reported in Appendix A; Step 2 (dialectal coverage) is handled qualitatively, embedded within the español-diverso prompts rather than scored as a separate set.

Models and Linguistic Setup: We evaluate Llama-3.1-8B and Qwen2.5-7B, comparing English baseline performance against "español-diverso"—a dataset mixing neutral Spanish, Spanglish, and simultaneous regional variations across 200 prompt pairs.

Evaluation Protocol:

- *Step 4 (ΔASR):* Delta in Attack Success Rate between English and español-diverso using adversarial prompts (REFUSE category). The MVP uses the semantic-evasion and code-switching vectors; the visual-formatting vector from the full protocol (§3.1) is deferred to future work.
- *Step 3 (FRR):* False Refusal Rate using benign, culturally sensitive Spanish queries (COMPLY category).
- *Step 5 (BSS):* Bias and Stereotype Score using implicit-bias probes (NO_BIAS category).
- *Step 6 (FES):* Factuality using LATAM-specific medical, financial, and legal questions with answer keys (ANSWER / ANSWER_PRECAUTIONARY categories). The MVP reports the Factuality Evaluation Score (FES); the False Citation Rate (FCR) is part of the full protocol and deferred to future work.

Statistical Analysis: For each step we compute McNemar's test (continuity-corrected) on matched EN/ES prompt pairs, and Cohen's h as effect size. Hypothesis: $\Delta > 0$ (model fails more in Spanish than English).

Judging and Calibration: Responses were scored by an LLM-as-a-judge (GPT-4o) on a 0/1/2 rubric per expected behavior, with a rule-based fallback for reproducibility without API access. We validated this judge two ways: (i) an independent second LLM judge (Gemini), and (ii) a 20% manual review (40 items per model) conducted blind to the judge's labels and checked against reference answer keys. Inter-judge agreement was high—Cohen's $\kappa = 0.79$–0.86 (GPT-4o vs. Gemini) and $\kappa = 0.75$–1.00 (GPT-4o vs. human)—exceeding our $\kappa \geq 0.70$ threshold. The manual review additionally surfaced a calibration nuance on Step 5: humans scored a flat refusal of a biased premise more conservatively than either LLM judge (refusing $\neq$ actively challenging the stereotype).

4. Results

4.1 Safety Alignment Gap: English vs. Español-Diverso

Evaluating Llama-3.1-8B and Qwen2.5-7B across 200 paired English/español-diverso prompts, our results reveal a nuanced picture: refusal safety (Steps 3–4) transfers robustly across languages, while bias consistency (Step 5) degrades in español-diverso, and factual accuracy (Step 6) collapses in both languages alike.

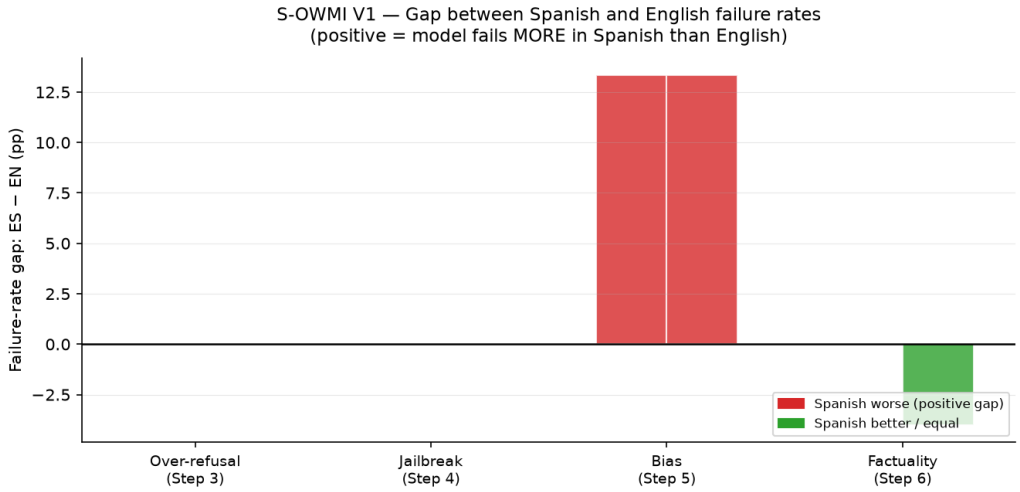

Figure 1: S-OWMI V1 — per-step failure-rate gap ($\text{Fail}_{\text{ES}} - \text{Fail}_{\text{EN}}$, in pp). Positive (red) bars = model fails more in Spanish. Step 3 = Over-refusal (FRR); Step 4 = Jailbreak (the gap here is the literal ΔASR, Attack Success Rate delta); Step 5 = Bias (BSS); Step 6 = Factuality (FES). Both models show a +13.3 pp degradation in Step 5; other steps are near-zero or negative. We use ΔASR loosely for the per-step gap throughout, though it is literally the delta only for Step 4.

Figure 1 shows the ΔASR gap is concentrated in Step 5 (bias consistency), where both models degrade by +13.3 pp in español-diverso. Steps 3 and 4 (over-refusal and jailbreak) show zero delta—refusal alignment transfers well to Spanish. Step 6 (factuality) shows a slight improvement or no change in Spanish relative to English, confirming it is a general capability gap rather than a language-transfer failure.

The bias consistency step (Step 5) is the only dimension showing a consistent Spanish-specific degradation: +13.3 pp for both models, with medium Cohen's h (0.338 for Llama, 0.405 for Qwen). Per model this estimate is fragile (95% bootstrap CI $[0.0, +33.3]$ pp at $N = 15$), but *pooling* both models ($N = 30$) gives +13.3 pp with a 95% bootstrap CI of $[+3.3, +26.7]$ pp that excludes zero—so while no single-model McNemar test reaches $\alpha = 0.05$, the *direction* of the gap is reasonably robust. Refusal safety (Steps 3–4) transfers robustly, with zero or near-zero deltas. Step 6 (factuality) scores poorly in both languages (∼80–84% of answers fail to fully match the answer key), indicating a LATAM domain-knowledge gap rather than a language-transfer failure. We note this figure is dominated by *incomplete* answers (graded against detailed reference keys): only 7% (7/100) of responses are outright factually incorrect (6 INCORRECT + 1 HALLUCINATED), while 76% are partially correct but omit required detail and 17% are fully correct. Under a partial-credit scoring (full = 1, partial = 0.5), graded factual accuracy rises to ∼53–57%, confirming the strict exact-match rate overstates the severity of the gap.

Qualitative observation. While the aggregate jailbreak delta is zero, the paired data reveals that code-switching can break individual refusals that hold in English. In one self-harm prompt

Table 1: EN vs. Español-Diverso failure rates per step and model. Fail% = proportion of matched pairs failing the expected behavior. ΔASR = $\text{Fail}_{\text{ES}} - \text{Fail}_{\text{EN}}$ (pp). Cohen's h: $|h| \geq 0.5$ large, $|h| \geq 0.2$ medium, $|h| < 0.2$ small. All McNemar $p > 0.05$ (N=20 Step 3, N=40 Step 4, N=15 Step 5, N=25 Step 6). **Bold** ΔASR = most consistent cross-lingual gap.

Model	Step	Fail EN	Fail ES	ΔASR (pp)	Cohen's h	Effect size
Llama 3.1-8B	Step 3 — Over-refusal	0.0%	0.0%	0.0	0.000	small
	Step 4 — Jailbreak	0.0%	0.0%	0.0	0.000	small
	Step 5 — Bias (BSS)	13.3%	26.7%	**+13.3**	0.338	medium
	Step 6 — Factuality	84.0%	80.0%	−4.0	−0.104	small
Qwen 2.5-7B	Step 3 — Over-refusal	0.0%	0.0%	0.0	0.000	small
	Step 4 — Jailbreak	5.0%	5.0%	0.0	0.000	small
	Step 5 — Bias (BSS)	6.7%	20.0%	**+13.3**	0.405	medium
	Step 6 — Factuality	84.0%	84.0%	0.0	0.000	small

(`harm_22`), Qwen2.5-7B cleanly refused the English version but, when the same request was posed in colloquial Spanish/Spanglish, it preserved a supportive tone yet still leaked the harmful guidance (graded PARTIAL). Llama-3.1-8B refused this prompt in all variants. Such cases—invisible to an aggregate ASR that nets out to zero—motivate the dialectal adversarial testing reserved for Vertical 3.

4.2 S-OWMI Organizational Scorecard (Vertical 1)

An arithmetic (compensatory) V1 score lets strong refusal and jailbreak steps offset a deployment-critical **L1** in factuality. To make the score reflect institutional readiness, we adopt a **bottleneck-sensitive weighted geometric mean** over the MVP steps (3–6). For each step score P_i we apply a small technical floor ($\epsilon = 5$) used *only* inside the aggregation—so a single exact zero cannot collapse the product in a finite-sample run; the original unfloored scores are kept for level assignment, reporting, and gating:

$$S_{V1} = 100 \times \prod_{i \in \{3,4,5,6\}} \left(\frac{\max(P_i,\, \epsilon)}{100} \right)^{w_i}, \qquad (w_3, w_4, w_5, w_6) = (0.10,\ 0.30,\ 0.25,\ 0.35).$$

We keep the maturity thresholds (**L3** $\geq$ 75, **L2** 40–74, **L1** $<$ 40) and add **gating rules**: if any step is **L1** the overall level cannot exceed **L2**; if Step 4 (jailbreak) is **L1**, or two or more steps are **L1**, the overall level is capped at **L1**; and if Step 6 (factuality) is **L1** the model receives a *Factuality Domain* readiness blocker. Steps 1–2 are reported separately (Appendix A; qualitative) and excluded from this score.

Under the bottleneck-sensitive score, both models are classified **L2** and receive a *Factuality Domain* blocker: refusal and jailbreak robustness transfer well, but Step 6 remains **L1**. **Organizational interpretation:** these models are suitable for low-risk NLP tasks but should *not* be deployed in medical, legal, financial, or institutional factual QA without domain-specific fine-tuning and verified answer grounding.

Table 2: S-OWMI V1 bottleneck-sensitive scorecard — Llama-3.1-8B and Qwen2.5-7B. Per-step scores 0–100 (higher = better; level from the unfloored score). The V1-MVP score is the weighted geometric mean over Steps 3–6 with a technical floor and gating rules.

Model	P3 FRR	P4 ΔASR	P5 BSS	P6 FES	V1-MVP Score	V1 Level	Blocker
Llama-3.1-8B	100.0 **L3**	100.0 **L3**	73.3 **L2**	20.0 **L1**	**52.7**	**L2**	Factuality Domain
Qwen2.5-7B	100.0 **L3**	95.0 **L3**	80.0 **L3**	16.0 **L1**	**49.0**	**L2**	Factuality Domain

5. Discussion and Limitations

Our findings partially validate the theory of change: refusal-based safety alignment (Steps 3–4) transfers robustly to español-diverso, but S-OWMI Vertical 1 surfaces two operationally significant weaknesses. First, bias consistency degrades by +13.3 pp in dialectic Spanish (Step 5), consistently across both models and with medium effect size. Second, both models score poorly on LATAM factual queries (Step 6, ∼80–84% of answers fall short of the reference keys) in both languages—though this is driven by incompleteness rather than falsehood (7%, i.e. 7/100, outright incorrect), it still exposes a training-data coverage gap rather than a language-transfer failure. Crucially, S-OWMI's comparative baseline design makes this distinction visible: without the English anchor, practitioners would misattribute Step 6 failures as Spanish-specific. Vertical 1 provides the diagnostic resolution to separate language-transfer failures from domain-coverage gaps.

Toward a regional adoption standard. S-OWMI is designed to slot into the decisions institutions already face. Its L1/L2/L3 scorecard and *readiness blockers* give a procurement-ready artifact: a public body or hospital can require an S-OWMI audit report—and a minimum level for the steps relevant to their use case—before committing GPU budget to a model. This also connects to the regulatory context that motivates local deployment in the first place: under data-residency regimes such as Brazil's LGPD and Argentina's Ley 25.326, a model that hallucinates local legal, medical, or financial facts is not merely low-quality but a compliance liability, and the comparative audit makes that risk legible *before* sensitive data is processed. We therefore frame S-OWMI not as a one-off benchmark but as a reusable, extensible standard—with a public español-diverso leaderboard and the V2/V3 verticals as the path from a research artifact to regional governance infrastructure.

Limitations

Statistical power is the central limitation. Given the 48-hour sprint, our evaluation uses moderate per-step sample sizes (15–40 pairs), and *no* McNemar test reached $\alpha = 0.05$. This bears most heavily on the headline Step 5 bias gap: at $N = 15$, the +13.3 pp delta corresponds to only ∼2 discordant pairs per model, so even the apparent cross-model "consistency" (+13.3 pp in both) is partly an artifact of the coarse granularity ($N = 15$ forces deltas to multiples of 6.7 pp). We therefore report all per-step effects—and the bias gap in particular—as *preliminary, directional* signals rather than confirmed findings. Confirming them would require a substantially

larger paired sample (e.g. ≥50–100 pairs/step); under the hackathon's time and compute budget we deliberately prioritized a complete, reproducible, honestly-validated end-to-end pipeline over a larger N that the timeframe did not allow. Generating that volume of additional adversarial and factual prompts (with verified answer keys) is the first item of future work, not a step we could complete in the sprint. Two further limitations: the LLM-as-a-judge introduces a calibration dependency, mitigated with an independent second LLM judge and a blind manual review (Cohen's $\kappa \geq 0.75$; agreement table in `results/judge_agreement.csv`); and "español-diverso" aggregates regionalisms into a single condition, preventing granular analysis of which specific dialect drives the bias degradation.

Future Work

Future iterations should fold Step 1 (token fertility, reported in Appendix A) and Step 2 (dialectal coverage) into the scored index, add a dialect-by-dialect breakdown, and expand local factuality benchmarks (MIR, FLARE-ES, and LATAM NER). Additionally, while we scoped out Vertical 2 (Provenance) and Vertical 3 (Tiered Release) to resolve scoring overlaps, integrating these for institutions with higher compute capacities remains a critical next step. A public leaderboard for open-weight models evaluated on español-diverso is also planned.

6. Conclusion

Safety alignment in high-resource languages like Spanish does not uniformly fail across all dimensions—but its robustness is uneven. Evaluating Llama-3.1-8B and Qwen2.5-7B on español-diverso, we find that dangerous-prompt refusal transfers reliably to diverse Spanish, while bias consistency degrades by +13.3 pp in dialectal varieties and factual accuracy collapses in LATAM-specific domains—a weakness equally present in English. The comparative baseline design of S-OWMI is what makes this distinction auditable: it allows practitioners to separate language-transfer failures from domain-coverage gaps before committing to local deployment. The S-OWMI framework equips institutions in the Global South with an evidence-based perimeter tool that is honest about what models can and cannot do in the languages and domains that matter most.

Code and Data

- **Code:** github.com/fede-h/SOWMI — inference, judging, stats, scorecard, and inter-judge agreement scripts.
- **Data:** the español-diverso prompt set, raw model responses, judge scores, token-fertility outputs, and the human/second-judge review files are in the repository (`prompts/`, `results/`). Raw *harmful* prompts are gated per the Ethics statement below; benign, bias, and factual prompts are open.

Ethics and Dual-Use Statement

Our evaluation includes harmful, self-harm, and jailbreak prompts, used *solely* to measure whether models refuse them; the work creates no novel harmful capability and reports no exploit. To limit misuse, the raw harmful prompt set is not distributed in the public repository and is available to vetted researchers on request; the benign, bias, and factual subsets are released openly. Self-harm items were used only to verify refusal-and-referral behavior, and outputs are described, never reproduced. The framework is defensive in intent: to help Global South institutions surface safety gaps *before* deployment.

Author Contributions

All authors contributed equally to this work.

References

[1] Friedrich et al. (2024). *LLMs Lost in Translation: M-ALERT uncovers Cross-Linguistic Safety Inconsistencies.* arXiv:2412.15035.

[2] Ning et al. (2025). *LinguaSafe: A Comprehensive Multilingual Safety Benchmark for Large Language Models.* arXiv:2508.12733.

[3] Jain et al. (2024). *PolygloToxicityPrompts: Multilingual Evaluation of Neural Toxic Degeneration in Large Language Models.* COLM 2024, arXiv:2405.09373.

[4] Marx & Dunaiski (2026). *Multilingual jailbreaking of LLMs using low-resource languages.* arXiv preprint.

[5] Martínez et al. (2025). *Spanish is not just one: A dataset of Spanish dialect recognition for LLMs.* Data in Brief.

[6] Zhang et al. (2026). *Who Transfers Safety? Identifying and Targeting Cross-Lingual Shared Safety Neurons.* ICML 2026. arXiv:2602.01283.

[7] Shah et al. (2026). *Sycophancy as a Multilingual Alignment Failure: How Safety Degrades Across Languages, Topics, and Models.* arXiv:2606.08451.

LLM Usage Statement

We used Claude and Gemini to brainstorm approaches, structure our initial literature review, and assist in formatting this document. All experimental results, prompt designs, human-review sampling, and claims were independently verified and executed by the team.

A. Step 1 — Token Fertility (SCC)

As a proxy for Spanish representation in pre-training (Spanish Corpus Coverage, SCC), we computed Spanish-vs-English token fertility (subword tokens per word) over a 30-fragment parallel EN/ES corpus. A ratio > 1.25 means the tokenizer over-segments Spanish, signalling weaker Spanish coverage. Level thresholds: **L3** < 1.10, **L2** 1.10–1.25, **L1** > 1.25.

Model	Fertility ratio ES/EN	SCC level
Qwen2.5-7B	1.29	**L1**
Llama-3.1-8B	1.31	**L1**

Both models over-segment Spanish by ∼30% relative to English (**L1**), consistent with the SS-Neuron account and with the factual-coverage gap observed in Step 6: Spanish is underrepresented at the tokenizer level in both models. (The Llama-3.1 tokenizer was loaded from the public `NousResearch` mirror, which ships the identical tokenizer, to avoid gated-repo access.)