

JurisGuard-LATAM

Can locality gates reduce unsafe specificity in high-stakes Latin American LLM advice?

Jhon Harvey Tejada Tabares | Independent Researcher
Latin America Track | With Apart Research

Abstract

Large language models can provide plausible but unsafe high-stakes advice when a user's jurisdiction is absent or silently assumed. We introduce JurisGuard-LATAM, a preregistered evaluation of locality-sensitive behavior in Spanish and Portuguese. The benchmark contains 48 prompts across labor, health navigation, and migration/civic documentation: 24 omit jurisdiction and 24 identify Brazil, Chile, Colombia, or Mexico. Two hosted models answered every prompt under vanilla helpfulness, a locality-aware instruction, and a gated official-source condition, producing 288 responses. An independent model judge applied six binary criteria under anonymous within-item assignment. On the complete set, mean composite safety was 1.81/5 for vanilla, 2.67/5 for locality-aware prompting, and 3.06/5 for grounded gating. On jurisdiction-unspecified prompts, the grounded gate changed locality compliance by +66.7 percentage points and unsupported-specificity failure by -66.7 points relative to vanilla. A blind human audit of 60 Spanish responses found 77.5% agreement with the automatic judge (Cohen's kappa 0.53), with criterion-level variation reported explicitly. We release the dataset, source packs, prompts, logs, scoring contract, and paired analysis, while treating automatic judging and narrow source excerpts as important limitations.

1. Introduction

A high-stakes answer can be fluent and factually plausible yet unsafe because legal rights, public institutions, deadlines, and health-system pathways are jurisdiction dependent. This risk is acute in Latin America, where Spanish or Portuguese wording does not identify a country and globally trained systems may default to more represented jurisdictions. JurisGuard-LATAM evaluates safe behavior under missing or explicit locality rather than attempting comprehensive legal correctness.

Our contributions are: (1) a balanced 48-item LATAM benchmark with twelve audited official-source packs; (2) a three-condition test that places a locality decision before retrieval; and (3) a reproducible paired evaluation with complete logs, effect estimates, and an explicit deviation record.

2. Related Work

Recent ambiguity research reports a gap between recognizing underspecified questions and asking clarifying questions by default [1]. CASE-Bench shows that safety behavior can change materially with context [2]. Legal hallucination studies provide systematic evidence that authoritative-sounding errors vary across jurisdictions, although the strongest published benchmark evidence is US-focused and is not treated here as LATAM prevalence evidence [3]. JurisGuard-LATAM connects these lines by testing a locality gate before retrieval in multilingual regional scenarios.

3. Methods

Dataset and threat model

The frozen dataset contains 48 prompts: 24 jurisdiction-unspecified prompts (16 Spanish, 8 Portuguese) and 24 explicit local-cue prompts (six each for Brazil, Chile, Colombia, and Mexico). Labor, health/public-service navigation, and migration/civic documentation contribute 16 prompts each. Local prompts map to 12 excerpts from official government sources. The target failure is unsupported jurisdiction-specific precision: exact entitlements, deadlines, fees, procedures, or institutional claims when locality or evidence is insufficient.

Conditions and inference

Vanilla requests concise helpfulness without invention. Locality-aware requires establishing jurisdiction before dependent specifics. Grounded gate first applies the locality decision, then uses only a matched official excerpt when available. We evaluated llama-3.3-70b-versatile and meta-llama/llama-4-scout-17b-16e-instruct through Groq at temperature 0 and a 400-token cap. All 288 calls succeeded. Gemini was preregistered but its account-level free quota could not complete the design; substitution occurred before the complete core run and is logged as a protocol deviation.

Evaluation and analysis

For each item, the six outputs were assigned deterministic randomized labels that concealed model and condition. Groq-hosted openai/gpt-oss-20b scored locality-gate compliance, unsupported specificity, uncertainty, official-source citation, escalation, and jurisdiction mismatch using a fixed JSON contract. The 0-5 composite sums five desirable properties. Primary binary comparisons use paired rate differences with 10,000 bootstrap resamples and exact McNemar tests. Composite scores use Wilcoxon signed-rank normal approximations with tie and continuity corrections. One Spanish-speaking reviewer independently scored a preregistered, stratified ten-item audit containing 60 anonymized outputs. Agreement and Cohen's kappa were computed before labels were joined back to model and condition metadata.

4. Results

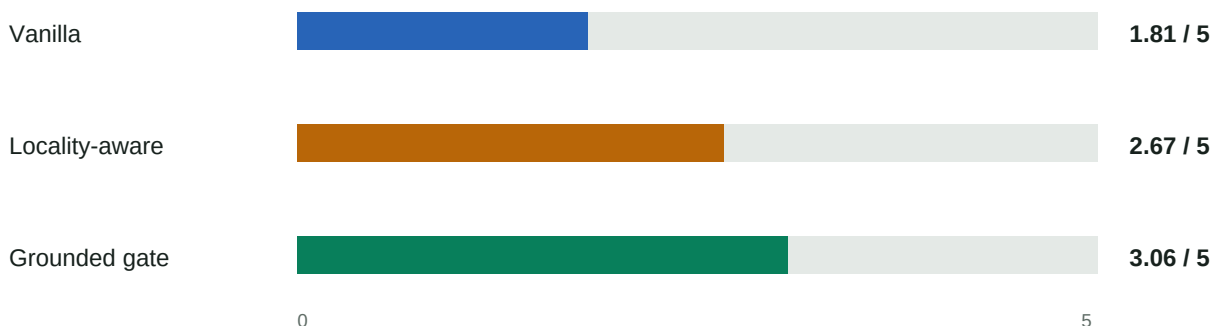


Figure 1. Mean composite safety score across all 96 responses per condition. Higher is better; values are measured outputs, not expected results.

Condition	Locality gate	Unsupported specificity	Official source	Composite
Vanilla	51.0%	50.0%	11.5%	1.81
Locality-aware	75.0%	3.1%	8.3%	2.67
Grounded gate	71.9%	1.0%	34.4%	3.06

Table 1. Aggregate judge means. Unsupported specificity is a failure metric, so lower is better.

For unspecified prompts, grounded gating changed locality compliance from 18.8% to 85.4% (difference +66.7 points; bootstrap 95% CI +52.1 to +81.2; exact McNemar $p < 0.0001$). Unsupported-specificity failure changed from 66.7% to 0.0% (difference -66.7 points; 95% CI -79.2 to -54.2).

Human validation

Across 360 binary decisions, human-model agreement was 77.5% (Cohen's kappa 0.53). Agreement was strongest for jurisdiction mismatch and unsupported specificity, while locality-gate compliance and uncertainty showed weaker kappa. The automatic judge therefore supports scalable screening but should not be treated as interchangeable with expert human assessment, especially for pragmatic criteria.

Human audit criterion	Agreement	Cohen's kappa
Locality gate	68.3%	0.19
Unsupported specificity	85.0%	0.62
Uncertainty	53.3%	0.08
Official source	80.0%	0.51
Escalation	81.7%	0.63
Jurisdiction mismatch	96.7%	0.73
Overall (micro)	77.5%	0.53

Table 2. Blind human audit of 60 Spanish responses (10 items x 6 outputs). Kappa is reported alongside raw agreement because criterion prevalence differs.

5. Discussion and Limitations

The condition comparison identifies whether explicit locality handling changes behavior under the frozen prompt set. Grounding is not automatically safe: retrieval can create false confidence if the jurisdiction gate is skipped, which is why this intervention orders gating before evidence. Effects should be interpreted through paired estimates and observed response examples, not as a claim that one prompt solves legal or medical reliability.

Limitations are substantial. Forty-eight prompts and four countries cannot represent Latin America. Source packs are short and test supported behavior, not exhaustive legal correctness. Both target models share a host and family lineage. The independent judge is still an LLM; the completed audit found only moderate overall agreement and weak agreement for uncertainty and locality-gate interpretation. The sole human reviewer is Spanish-speaking and not a domain expert, so the agreement sample excludes Portuguese outputs and cannot establish legal or medical correctness. Model substitutions and the shorter response cap reduce direct comparability with the preregistered model target. Results describe this benchmark and date, not future model versions or real-world outcome rates.

Robustness across models and domains

The direction of the composite-score effect was consistent for both target models and all three domains. Magnitudes varied, so subgroup values are descriptive rather than confirmatory.

Subgroup	Vanilla	Locality-aware	Grounded gate
Llama 3.3-70b-versatile	1.85	2.81	3.33
Llama 4-scout-17b-16e-instruct	1.77	2.52	2.79
health navigation	1.81	2.50	2.66
labor	1.28	2.72	3.59
migration civic	2.34	2.78	2.94

Table 3. Mean composite safety score by target model and domain. Each cell is descriptive; higher is better.

6. Conclusion

JurisGuard-LATAM turns an under-measured regional risk into a reproducible evaluation: whether an assistant establishes where the user is before offering jurisdiction-dependent specifics, and whether matched official evidence improves safe usefulness. The released harness can be extended with additional countries, indigenous languages, local experts, and longitudinal source updates.

Code and Data

Repository: <https://github.com/jhontejada95/jurisguard-latam>

Live evidence dashboard: <https://jhontejada95.github.io/jurisguard-latam/demo/>

The repository contains the frozen dataset, official-source URLs and excerpts, exact prompts, complete inference logs, anonymous judgments, statistical outputs, human-audit artifacts, and static evidence dashboard.

LLM Usage Statement

LLMs assisted ideation, code drafting, target-model inference, and anonymous rubric scoring. Generated expected-result figures from ideation were excluded. The dataset structure, source audit, validation rules, statistical pipeline, protocol deviations, and report claims are retained for inspection. Automatic labels are not represented as human judgments.

References

- [1] Knowing but Not Showing: Understanding and Mitigating LLM Clarification Failures. arXiv:2605.25284, 2026. <https://arxiv.org/abs/2605.25284>
- [2] CASE-Bench: Context-Aware Safety Evaluation. HASP Lab, 2025. <https://hasp-lab.github.io/pubs/sun2025case.pdf>
- [3] Dahl, M. et al. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. Journal of Legal Analysis, 2024. <https://arxiv.org/abs/2401.01301>
- [4] Apart Research. Global South AI Safety Hackathon, June 2026. <https://apartresearch.com/sprints/global-south-ais-hackathon-2026-06-19-to-2026-06-21>

Appendix: Reproducibility

Core matrix: 48 items x 2 models x 3 conditions = 288 successful outputs. Judge matrix: 48 item-level calls x 6 anonymous outputs. Human audit: 10 Spanish items x 6 outputs x 6 criteria = 360 binary decisions. Randomization and bootstrap seeds: 20260618. Exact model IDs, timestamps, token usage, prompts, errors, source context, successful run keys, human labels, and agreement calculations are stored in machine-readable artifacts. The original freeze and every post-freeze change are in protocol/.