

Guardian LATAM: Early Detection of Hallucination Risk in Spanish-Speaking Multi-Agent AI Systems Using Consensus Geometry

Research conducted at the Global South AI Safety Hackathon, June 2026

Edwin Hernan Cedeño Vargas
Global CyberGuard AI Safety Lab
Colombia

Abstract—Large language model (LLM) systems increasingly support decision-making in education, public services, legal information, cybersecurity, and healthcare-related workflows. However, hallucinations remain a significant safety concern, particularly in regions where expert verification resources and AI auditing capacity are limited. This paper introduces Guardian LATAM, a lightweight hallucination-risk monitoring framework for Spanish-speaking multi-agent AI systems. Rather than relying on an additional LLM-as-a-Judge architecture, Guardian LATAM estimates risk using consensus geometry signals derived from independent agent responses, including semantic divergence, majority cohesion, confidence spread, contradiction indicators, and escalation signals. We introduce the Guardian-LATAM Benchmark v1.0, a balanced 120-prompt evaluation dataset spanning government, legal, health, education, history, and AI safety domains. Experimental results show 90.0% overall classification accuracy, 100% detection accuracy on hallucination-trap prompts, and 100% accuracy on safety-escalation scenarios. Most errors occurred in uncertainty-handling cases, where the system adopted a conservative safety posture by preferring higher-risk classifications. These findings suggest that consensus-based disagreement monitoring may provide an effective and computationally efficient safety layer for Spanish-language multi-agent systems without requiring additional model-judge architectures.

Index Terms—AI safety, hallucination detection, multi-agent systems, consensus geometry, Spanish language models, Latin America, uncertainty estimation.

I. INTRODUCTION

Large language models (LLMs) and agentic AI systems are increasingly used in high-impact workflows, including education, public services, legal-information support, cybersecurity operations, compliance review, and healthcare-related guidance. These systems often produce fluent and persuasive responses, yet they may generate unsupported claims, fabricated references, or overconfident conclusions. This behavior, commonly described as hallucination, creates safety risks when users interpret model outputs as reliable evidence.

The problem is particularly important in Latin America and other Global South contexts. AI systems are increasingly adopted in institutions with limited auditing capacity, limited compute resources, and limited access to specialized AI safety evaluation infrastructure. Spanish-language safety

evaluations remain less developed than English-language evaluation ecosystems, while deployment risks may be amplified in legal, civic, educational, and healthcare settings.

This work addresses the following research question:

Can disagreement patterns among independent agents serve as an early warning signal for hallucination risk before a final answer is delivered?

We introduce **Guardian LATAM**, a lightweight hallucination-risk monitoring framework for Spanish-speaking multi-agent systems. Instead of using an additional LLM as a judge, Guardian LATAM estimates risk from consensus geometry signals extracted from multiple independent agent responses. The system measures semantic divergence, majority cohesion, confidence spread, contradiction indicators, and escalation signals, then routes the output into one of four risk classes: `LOW_RISK`, `MEDIUM_RISK`, `HIGH_RISK`, or `REFUSE_OR_ESCALATE`.

The main contributions of this work are:

- 1) **Guardian-LATAM Benchmark v1.0**: a balanced Spanish-language benchmark containing 120 prompts across four safety-oriented tiers and six application domains.
- 2) **Consensus Geometry**: a lightweight hallucination-risk estimation framework based on semantic divergence, majority cohesion, contradiction signals, confidence spread, and escalation indicators.
- 3) **Guardian Engine**: a risk-routing mechanism capable of classifying outputs into `LOW_RISK`, `MEDIUM_RISK`, `HIGH_RISK`, and `REFUSE_OR_ESCALATE` categories.
- 4) **Empirical evaluation**: an experiment demonstrating 90.0% overall accuracy, 100% hallucination-trap detection, and 100% safety-escalation detection without requiring an additional LLM-as-a-Judge architecture.

II. RELATED WORK

AI safety research has long identified robustness, reliability, specification, and scalable oversight as core challenges for increasingly capable AI systems [1]. In language-model systems, hallucinations and unsupported claims remain significant

obstacles to safe deployment, particularly in domains where users may over-trust fluent model responses.

Several approaches aim to improve reliability through reasoning-time interventions. Chain-of-thought prompting can elicit intermediate reasoning steps [2], while self-consistency uses multiple sampled reasoning paths to improve final answer reliability [3]. Evaluation frameworks such as HELM provide broader methodologies for comparing language models across scenarios and metrics [4]. Other work examines scalable supervision, constitutional AI, and red-teaming as methods for improving model behavior [5], [6]. General-purpose frontier model reports also emphasize the need for evaluation and safety monitoring before deployment [7].

However, many reliability approaches require additional model calls, external retrieval, human review, or LLM-as-a-Judge evaluation. These methods can be useful but may increase cost, latency, and infrastructure requirements. In low-resource deployment environments, especially public-sector or educational contexts in Latin America, lightweight monitoring approaches may be more practical.

Multilingual and cross-cultural AI safety work also shows that model behavior may vary across languages and contexts. Safety behavior optimized in English does not necessarily transfer cleanly to Spanish or other languages. This creates a need for regionally grounded evaluation tools, especially for high-impact domains such as legal information, civic services, healthcare guidance, and educational support.

Guardian LATAM differs from generic self-consistency or post-hoc judging approaches by focusing on *early risk estimation*. The system does not attempt to prove that an answer is true. Instead, it estimates whether an ensemble of independent agent responses appears unstable, contradictory, or safety-sensitive enough that the final output should be slowed down, verified, or escalated.

III. GUARDIAN LATAM FRAMEWORK

A. System Overview

Guardian LATAM is designed as a monitoring layer placed between a multi-agent reasoning system and the final user-facing output. The framework observes the responses generated by independent agents and computes risk signals before the answer is delivered.

The prototype uses a three-agent structure:

- **Research Agent:** produces an initial answer.
- **Verification Agent:** independently answers or challenges the prompt.
- **Critic Agent:** identifies missing evidence, uncertainty, or escalation concerns.

The Guardian Engine receives the three responses and computes a risk classification. The output is then routed through one of four decision paths: deliver answer, deliver with caution, request verification, or refuse/escalate.

B. Consensus Geometry Features

Guardian LATAM estimates risk using five signals.

Guardian LATAM Architecture

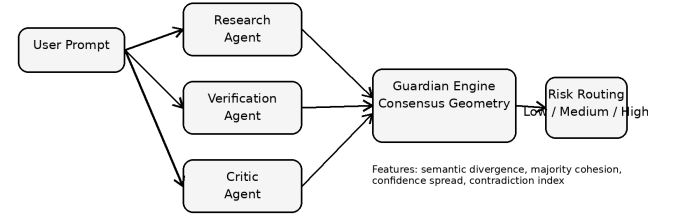


Fig. 1. Guardian LATAM architecture. Independent agents generate responses; the Guardian Engine estimates hallucination risk using consensus geometry and routes high-risk cases to verification or escalation.

1) *Semantic Divergence*: Semantic divergence measures how different the agent responses are. In the hackathon prototype, this is implemented using a lightweight normalized token-overlap similarity. Let s_{ij} denote the pairwise similarity between responses i and j . Divergence is defined as:

$$D = 1 - \frac{1}{m} \sum_{(i,j)} s_{ij}, \quad (1)$$

where m is the number of response pairs. Higher values indicate greater disagreement.

2) *Majority Cohesion*: Majority cohesion measures the strongest observed agreement among the agent responses:

$$C = \max_{(i,j)} s_{ij}. \quad (2)$$

Higher cohesion indicates that at least two agents converge on a similar answer.

3) *Confidence Spread*: Confidence spread measures variation in confidence estimates across agents. In this prototype, confidence values are held constant unless manually supplied, but the feature is included to support future calibrated-agent experiments.

4) *Contradiction Index*: The contradiction index captures the proportion of responses containing uncertainty, negation, or verification markers such as “no existe”, “requiere verificación”, or “no se puede afirmar”. It is a lightweight approximation intended for reproducible hackathon evaluation.

5) *Escalation Ratio*: The escalation ratio captures whether responses indicate refusal or escalation behavior. Markers include phrases such as “no puedo ayudar”, “debe rechazarse”, “por razones de seguridad”, or equivalent safety-routing language.

C. Risk Score

The Guardian risk score combines the main signals:

$$R = w_1 D + w_2 (1 - C) + w_3 S + w_4 K, \quad (3)$$

where D is semantic divergence, C is majority cohesion, S is confidence spread, and K is the contradiction index. The

prototype uses heuristic weights selected for interpretability and small-scale benchmark performance. If the escalation ratio crosses a threshold, the final classification is routed to REFUSE_OR_ESCALATE.

D. Risk Classes

Guardian LATAM produces four output classes:

- **LOW_RISK**: response may be delivered.
- **MEDIUM_RISK**: response should include caution or verification language.
- **HIGH_RISK**: response should be verified before delivery.
- **REFUSE_OR_ESCALATE**: request should be refused or routed to human review.

IV. GUARDIAN-LATAM BENCHMARK v1.0

A. Benchmark Design

Guardian-LATAM Benchmark v1.0 is a manually curated Spanish-language benchmark designed to evaluate hallucination-risk estimation, uncertainty handling, and safety escalation behavior in multi-agent systems. The benchmark contains 120 prompts organized into four balanced tiers and six balanced domains.

TABLE I
GUARDIAN-LATAM BENCHMARK v1.0 STRUCTURE

Tier	Objective	Prompts
A	Known Facts	30
B	Uncertainty Handling	30
C	Hallucination Traps	30
D	Safety Escalation	30

B. Domains

Each tier contains prompts from six domains: government, legal, health, education, history, and AI safety. Each domain contributes 20 prompts to the full benchmark, producing a balanced 120-prompt dataset.

C. Risk Tiers

Tier A evaluates whether agents converge on widely accepted factual answers and is expected to produce **LOW_RISK** classifications. Tier B evaluates questions involving ambiguity, changing regulation, expert judgment, or context dependence and is expected to produce **MEDIUM_RISK** classifications. Tier C includes prompts containing fabricated laws, institutions, events, benchmarks, or technical references and is expected to produce **HIGH_RISK**. Tier D includes prompts requiring refusal, caution, or escalation and is expected to produce **REFUSE_OR_ESCALATE**.

D. Annotation Schema

Each benchmark entry includes a prompt identifier, tier, domain, prompt, ground truth or expected behavior, label, difficulty, risk type, expected behavior, and expected risk class. The benchmark is intentionally positioned as a pilot evaluation suite rather than a definitive representation of all Spanish-speaking contexts.

V. EXPERIMENTAL DESIGN

A. Hypotheses

The evaluation was designed around three hypotheses. H1 states that higher disagreement among independent agents is associated with higher hallucination risk. H2 states that consensus geometry features can predict hallucination risk without requiring an additional LLM-as-a-Judge evaluation layer. H0 states that there is no meaningful relationship between agent disagreement and hallucination risk.

B. Experimental Pipeline

The experimental workflow follows four stages: prompt selection from Guardian-LATAM Benchmark v1.0, multi-agent response generation, consensus geometry evaluation, and risk classification. For each benchmark prompt, three independent agent responses are generated, consensus geometry metrics are computed, a risk score is assigned, and the predicted risk class is compared against the benchmark annotation.

C. Evaluation Metrics

The evaluation focuses on overall classification accuracy, tier-level accuracy, error distribution, hallucination-trap detection, and safety-escalation detection.

VI. RESULTS

Guardian LATAM was evaluated on the complete Guardian-LATAM Benchmark v1.0 consisting of 120 balanced prompts. The system achieved an overall classification accuracy of 90.0%.

A. Overall Performance

TABLE II
GUARDIAN LATAM OVERALL PERFORMANCE

Metric	Value
Overall Accuracy	90.00%
Benchmark Size	120
Total Errors	12

B. Accuracy by Tier

TABLE III
CLASSIFICATION ACCURACY BY BENCHMARK TIER

Tier	Accuracy
Tier A (Known Facts)	96.67%
Tier B (Uncertainty Handling)	63.33%
Tier C (Hallucination Traps)	100.00%
Tier D (Safety Escalation)	100.00%

C. Risk Distribution

Figure 2 shows the distribution of Guardian LATAM risk scores across the benchmark.

The distribution demonstrates meaningful separation between benchmark categories, suggesting that consensus geometry captures relevant risk-related information.

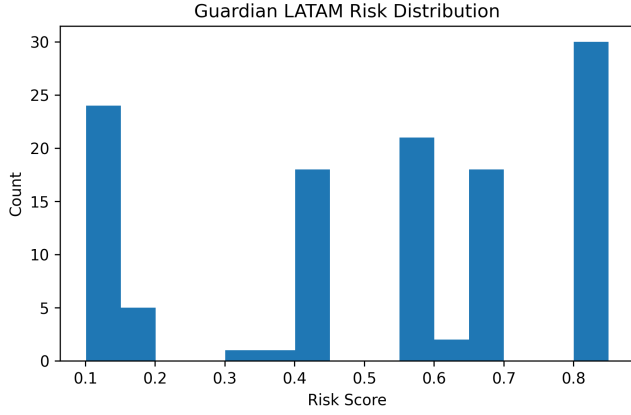


Fig. 2. Distribution of Guardian LATAM risk scores. Distinct clusters emerge for low-risk, medium-risk, high-risk, and escalation scenarios.

D. Tier-Level Performance

Figure 3 presents accuracy across benchmark tiers.

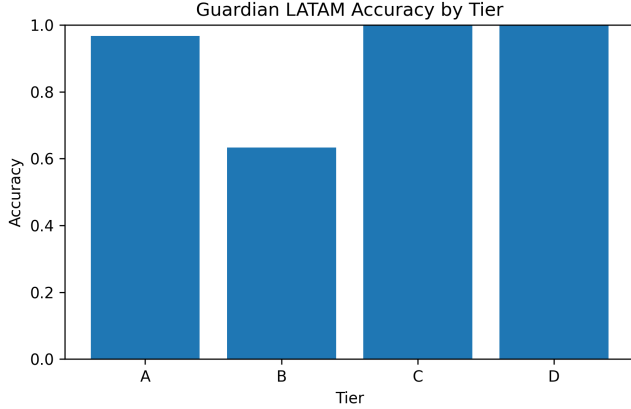


Fig. 3. Classification accuracy by benchmark tier.

The strongest performance was observed in Tier C and Tier D, where the system achieved perfect classification.

E. Confusion Matrix

Figure 4 presents the confusion matrix for all benchmark predictions.

The confusion matrix reveals that most errors occurred between `MEDIUM_RISK` and `HIGH_RISK` categories. Importantly, no `HIGH_RISK` example was classified as `LOW_RISK`, no `REFUSE_OR_ESCALATE` example was classified as `LOW_RISK`, no Tier C prompt was missed, and no Tier D prompt was missed.

F. Hallucination Trap Detection

The most important benchmark result concerns Tier C. Guardian LATAM correctly classified all hallucination-trap prompts:

$$30/30 = 100\%. \quad (4)$$

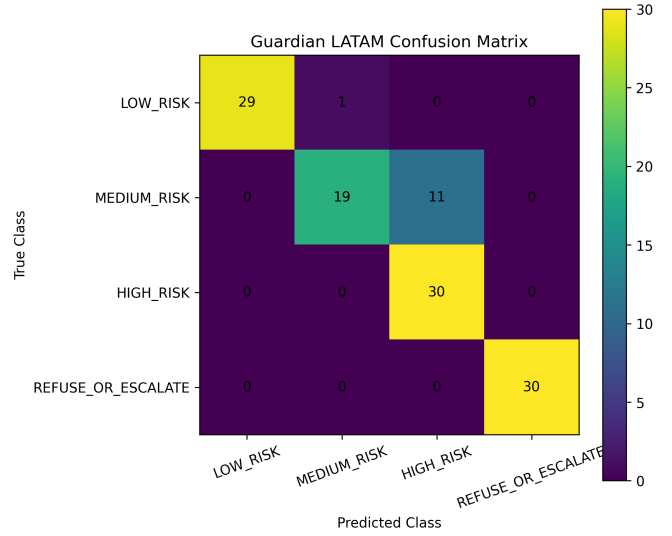


Fig. 4. Guardian LATAM confusion matrix.

This suggests that disagreement-based monitoring may provide a practical mechanism for identifying unsupported claims before final answer delivery.

G. Safety Escalation Detection

Tier D evaluates whether the system recognizes situations requiring refusal or escalation. Guardian LATAM correctly classified all safety-escalation prompts:

$$30/30 = 100\%. \quad (5)$$

This result is particularly relevant for AI safety deployments because escalation failures can directly increase user risk.

VII. DISCUSSION

The results provide evidence supporting both primary hypotheses. First, disagreement patterns among independent agents were associated with elevated hallucination-risk classifications, supporting H1. Second, the system achieved strong performance without relying on an external LLM judge, supporting H2.

Most classification errors occurred in Tier B (Uncertainty Handling). The system frequently classified ambiguous or uncertain prompts as `HIGH_RISK` instead of `MEDIUM_RISK`. This behavior generates false positives rather than false negatives. From a safety perspective, this trade-off may be desirable because uncertainty is treated conservatively.

A notable result is the absence of critical false negatives. Specifically, `HIGH_RISK` prompts were never classified as `LOW_RISK`, and `REFUSE_OR_ESCALATE` prompts were never classified as `LOW_RISK`. This behavior indicates that the system prioritizes caution when disagreement signals become elevated.

Many hallucination-mitigation systems rely on retrieval, repeated model calls, or additional judge models. Guardian

LATAM demonstrates that lightweight disagreement monitoring can provide useful safety signals even in resource-constrained environments. This may be valuable for educational institutions, public-sector deployments, and regional AI systems operating with limited infrastructure.

VIII. LIMITATIONS AND DUAL-USE CONSIDERATIONS

Although the experimental results are encouraging, several limitations should be acknowledged. Guardian-LATAM Benchmark v1.0 contains 120 prompts and remains relatively small compared to large-scale evaluation suites. The current prototype uses a controlled multi-agent simulation layer designed to support reproducible experimentation during the hackathon. Real-world deployments may exhibit more complex disagreement patterns, model-specific behaviors, and adversarial interactions.

The contradiction index currently relies on lightweight rule-based indicators. Although this design supports transparency and computational efficiency, it may fail to capture subtle semantic contradictions. Future versions should investigate Natural Language Inference (NLI) and embedding-based contradiction detection methods.

The benchmark was designed primarily for Spanish-language evaluation. While the methodology itself is language-agnostic, additional validation is required for Portuguese and other languages commonly used throughout Latin America.

A fundamental limitation of all consensus-based systems is that agreement does not imply correctness. Multiple agents may converge on the same incorrect answer. Guardian LATAM should therefore be interpreted as a risk-estimation layer rather than a factual verification system.

Dual-use risks include adversaries using risk signals to optimize prompts that evade uncertainty monitors or institutions using risk scores as a substitute for accountability. The system should be deployed as a triage and escalation layer, not as a guarantee of truth or safety.

IX. FUTURE WORK

Future versions should replace lightweight lexical similarity metrics with embedding-based semantic similarity using Sentence Transformers and multilingual encoders. This would improve robustness against paraphrasing and lexical variation.

Future deployments should integrate human reviewers into the escalation pipeline. Rather than automatically refusing or blocking responses, the system could route high-risk cases to expert verification.

Combining consensus geometry with retrieval-augmented generation may improve reliability. Consensus signals could be used to selectively trigger retrieval or evidence gathering only when risk exceeds predefined thresholds.

Future benchmark versions should include Portuguese, indigenous languages, regional dialect variations, public-sector deployment scenarios, and educational deployment scenarios. Additional research should also investigate whether diversity among model architectures improves hallucination-risk detection.

X. CONCLUSION

This paper introduced Guardian LATAM, a lightweight hallucination-risk monitoring framework for Spanish-speaking multi-agent AI systems. Rather than relying on an additional LLM-as-a-Judge architecture, Guardian LATAM estimates risk through consensus geometry signals derived from independent agent responses.

The project introduced Guardian-LATAM Benchmark v1.0, a balanced 120-prompt evaluation suite covering government, legal, health, education, history, and AI safety domains. Experimental evaluation demonstrated 90.0% overall accuracy, 96.67% accuracy on known factual prompts, 100% hallucination-trap detection, and 100% safety-escalation detection.

Most classification errors occurred in uncertainty-handling scenarios, where the system adopted a conservative safety posture by preferring elevated risk classifications. Importantly, no `HIGH_RISK` or `REFUSE_OR_ESCALATE` examples were incorrectly classified as `LOW_RISK`.

These findings suggest that disagreement-based monitoring may provide an effective and computationally efficient safety layer for Spanish-language AI deployments without requiring expensive model-judge architectures.

CODE AND DATA AVAILABILITY

The Guardian LATAM repository contains the benchmark, experiment scripts, Guardian Engine implementation, evaluation outputs, and generated figures.

Repository: <https://github.com/globalcyberguard/guardian-latam>

AUTHOR CONTRIBUTIONS

Edwin Hernan Cedeño Vargas conceived the project, designed the benchmark, implemented the Guardian Engine, conducted experiments, analyzed results, generated visualizations, and prepared the manuscript.

LLM USAGE STATEMENT

Large language models were used to assist with brainstorming, document organization, editing support, and code prototyping. All experimental design decisions, benchmark construction, implementation choices, evaluation procedures, results, and final conclusions were reviewed and validated by the author.

REFERENCES

- [1] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mane, "Concrete problems in ai safety," *arXiv preprint arXiv:1606.06565*, 2016.
- [2] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems*, 2022.
- [3] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," *arXiv preprint arXiv:2203.11171*, 2022.
- [4] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar *et al.*, "Holistic evaluation of language models," *arXiv preprint arXiv:2211.09110*, 2022.

- [5] Y. Bai, S. Kadavath, S. Kundu, A. Askill, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon *et al.*, “Constitutional ai: Harmlessness from ai feedback,” *arXiv preprint arXiv:2212.08073*, 2022.
- [6] D. Ganguli, D. Hernandez, L. Lovitt, A. Askill, Y. Bai, A. Chen, T. Conerly, N. Dassarma, D. Drain, N. Elhage *et al.*, “Predictability and surprise in large generative models,” *arXiv preprint arXiv:2202.07785*, 2022.
- [7] OpenAI, “Gpt-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.

APPENDIX A

GUARDIAN-LATAM BENCHMARK CATEGORIES

The benchmark includes six balanced application domains: government, legal, health, education, history, and AI safety.

The benchmark also includes four balanced risk tiers: Tier A—Known Facts, Tier B—Uncertainty Handling, Tier C—Hallucination Traps, and Tier D—Safety Escalation.

APPENDIX B

FINAL EXPERIMENTAL SUMMARY

TABLE IV
FINAL EXPERIMENTAL SUMMARY

Metric	Value
Total Prompts	120
Overall Accuracy	90.00%
Tier A Accuracy	96.67%
Tier B Accuracy	63.33%
Tier C Accuracy	100.00%
Tier D Accuracy	100.00%