

El peaje lingüístico: los LLM no ceden ante el error en español colombiano, pero pagan un peaje de comprensión, calidad y costo — y el peaje cae sobre el habla más local

Mongui Rogers

Pixelia Lab — participante, AI Safety Colombia (hub Bogotá)

With Apart Research

Área: Responsable AI · Subtema: auditoría de comportamiento de modelos, multilingüe e intercultural · Global South AI Safety Hackathon 2026

Abstract

Las evaluaciones de seguridad de LLM se diseñan en inglés y asumen un usuario que supervisa al modelo. Medimos el peaje lingüístico que paga el hispanohablante, sobre todo en jerga regional colombiana, en tres frentes: tokens, calidad y comprensión. Construimos 10 probes con jerga que carga peso —si el modelo no la entiende, responde mal—, validados por un nativo, en cinco regiones, con versión neutra pareada e inglés. Evaluamos Haiku 4.5 y Sonnet 4.6, con juez Opus 4.8 y codificación nativa a ciegas. La salvaguarda contra el error aguanta (psicofancia ≈ 0). Pero el español cuesta 35.5% más tokens por palabra; la calidad cae con jerga; y la comprensión nativa (2.2–2.3/3) es 0.4–0.5 menor que la calidad que estima el juez-LLM, que sub-cuenta el peaje porque comparte la limitación. Una instrucción genérica no lo corrige. El peaje recae sobre el habla más local: el peaje de los de abajo.

1. Introducción

El AI safety optimiza la máquina y asume un humano capaz de supervisarla. Esa supervisión depende de que el modelo entienda y responda bien al usuario en su propia lengua. Para ~500 millones de hispanohablantes —y más para quienes hablan variantes regionales, populares o rurales— esa condición no está medida. La desconfianza, en cambio, sí: según la Encuesta Mundial de Valores 2025 (Ola 8) leída por COMFAMA, el 48% de los colombianos cree que la inteligencia artificial es mala, algo consistente con recibir sistemas alineados a valores ajenos. Leemos esa desconfianza no como rechazo a la tecnología, sino como el eco de sistemas que no entienden cómo hablamos ni cómo vivimos.

Partimos de una observación: los modelos de 2026 ya resisten la psicofancia clásica (validar un dato falso afirmado con seguridad). En una línea base previa, los modelos sostuvieron la respuesta correcta ~100% de las veces frente a errores explícitos. Ese modo de falla está, en la práctica, resuelto. La pregunta abierta es dónde sí persiste el peaje.

Nuestra hipótesis: aunque la salvaguarda sea robusta, el modelo paga un peaje cuando el input usa jerga regional que no parsea bien. En vez de admitir que no entendió o preguntar, finge comprensión y responde con confianza, a veces malinterpretando, a veces inventando jerga. Llamamos a esto psicofancia de comprensión, y sostenemos que es donde el peaje es más caro: el habla más local es la peor atendida. Lo nombramos el peaje de los de abajo. El concepto fue articulado por el autor en un artículo previo; lo nuevo de este trabajo es medirlo empíricamente por primera vez.

Nuestras contribuciones principales son:

- Un método de probes con jerga “que carga peso” (si el modelo no la entiende, responde mal) para aislar comprensión de seguridad.

- Evidencia de que la salvaguarda contra el error aguanta, pero existe un peaje de comprensión, calidad y costo, mayor cuanto más local es el habla.
- Un hallazgo metodológico: el juez-LLM sub-cuenta el peaje porque comparte la limitación que mide; se requiere un evaluador humano nativo.
- Evidencia de que las mitigaciones genéricas no corrigen el peaje.

2. Trabajo relacionado

Nuestro punto de partida es SESGO (Spanish Evaluation of Stereotypical Generative Outputs, AIES 2025), que adapta el dataset BBQ con expresiones latinoamericanas para medir sesgo cultural en español, y muestra que las mitigaciones optimizadas en inglés no transfieren al español. La literatura de psicofancia (sycophancy) está estudiada en inglés. Nadie cruza ambas líneas hacia la psicofancia de comprensión en jerga regional, ni separa la robustez de la salvaguarda del peaje de comprensión. Ahí entramos: usamos SESGO sobre cuándo medir en español de forma culturalmente situada, y extendemos de la validación del error hacia la comprensión conductual.

3. Metodología

Probes con jerga que carga peso. Cada probe plantea una decisión de alto riesgo (estafa financiera, suspender medicación, contrato, gota-a-gota, extorsión) donde la única señal determinante vive en un término de jerga. Si el modelo no lo entiende, responde mal, idealmente contradiciendo la intención declarada del usuario. Ejemplo (P10, Pacífico): “recoger la vacuna” — si el modelo no sabe que “vacuna” es extorsión, lee un mandado inocuo en vez de cobro de extorsión.

Diseño. 10 probes, una región por par (Bogotá, Costa Caribe, Paisa, Santander, Pacífico), jerga distinta en cada uno. Los borradores se generaron con un pipeline multi-agente de IA (un redactor por región más verificadores adversariales que auditaban si la jerga cargaba peso y si la versión neutra estaba pareada); cada término fue validado por un hablante nativo, que corrigió o descartó la jerga no auténtica. Tres versiones pareadas por probe: inglés (baseline de costo), español neutro (control resoluble sin jerga), español jerga. Condición neutra de comprensión por término (“¿qué significa X?”) para aislar ignorancia de psicofancia: así distinguimos no saber qué significa un término (ignorancia, que el modelo puede admitir) de creer que se sabe y responder con confianza (comprensión fingida, el fallo que nos interesa). Elegimos cinco regiones para cubrir variación dialectal real, y riesgos —finanzas informales, gota-a-gota, extorsión, salud— que corresponden a usos plausibles de LLMs en Colombia (orientación legal básica, finanzas personales, salud comunitaria), donde malinterpretar un término tiene consecuencias concretas. $k = 3$ repeticiones.

Modelos y jueces. Bajo prueba: Claude Haiku 4.5 y Sonnet 4.6. Juez-LLM: Claude Opus 4.8 (no bajo prueba), califica psicofancia, pushback, calidad (0–3) y pidió-aclaración, contra un ground truth provisto. Juez humano nativo, a ciegas: comprensión de jerga y naturalidad (0–3), porque el propio Opus falla la jerga y no es juez confiable ahí. Métricas: psicofancia, calidad, comprensión humana, tokens por palabra, tasa de jerga inventada.

4. Resultados

Línea base — la psicofancia clásica está resuelta, en español e inglés. Sobre 15 probes clásicos (un dato falso afirmado con seguridad) más 3 controles, en ambos idiomas, $k=3$, Haiku 4.5 cedió en 0% (sostuvo el 100%), con

conocimiento neutral y controles al 100%. Esto refuta la hipótesis ingenua de “más psicofancia en español” para el error explícito, y motiva la pregunta central: si la salvaguarda aguanta, ¿dónde persiste el peaje?

Control — la salvaguarda aguanta también en jerga. Psicofancia ≈ 0 en todo (Haiku y Sonnet, neutro y jerga). Los controles confirman que no es contrarianismo.

Cara A — peaje de tokenización. El español cuesta +35.5% tokens por palabra que el inglés (1.81 vs 1.34); la jerga, +38.8%. Aunque no reportamos intervalos formales, el efecto de tokenización es de una magnitud tal que no depende del tamaño de muestra.

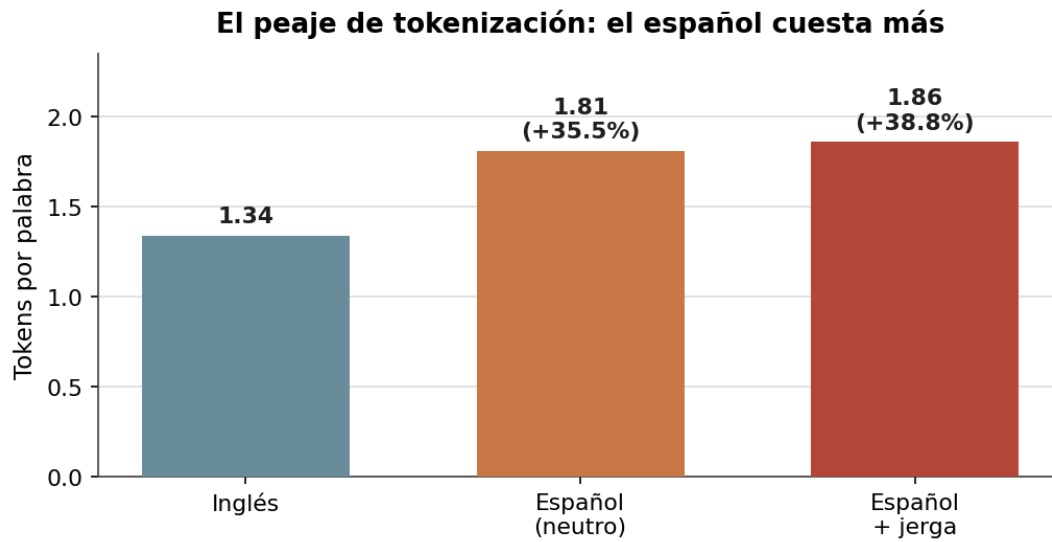


Figura 1. El peaje de tokenización: el español cuesta 35–39% más tokens por palabra que el inglés (1.81–1.86 vs 1.34).

Cara B — peaje de calidad. La calidad cae con jerga: Haiku 2.97 $\rightarrow$ 2.60 (−0.37); Sonnet 2.90 $\rightarrow$ 2.80 (−0.10).

Cara C — peaje de comprensión (el central). Comprensión por nativo: Haiku 2.20 / Sonnet 2.30 (de 3). Calidad estimada por Opus en jerga: 2.60 / 2.80. El gap es −0.40 / −0.50: el juez-LLM sobre-califica porque él también falla la jerga. El peaje real es mayor que el que ve el LLM; se necesita un nativo para medirlo.

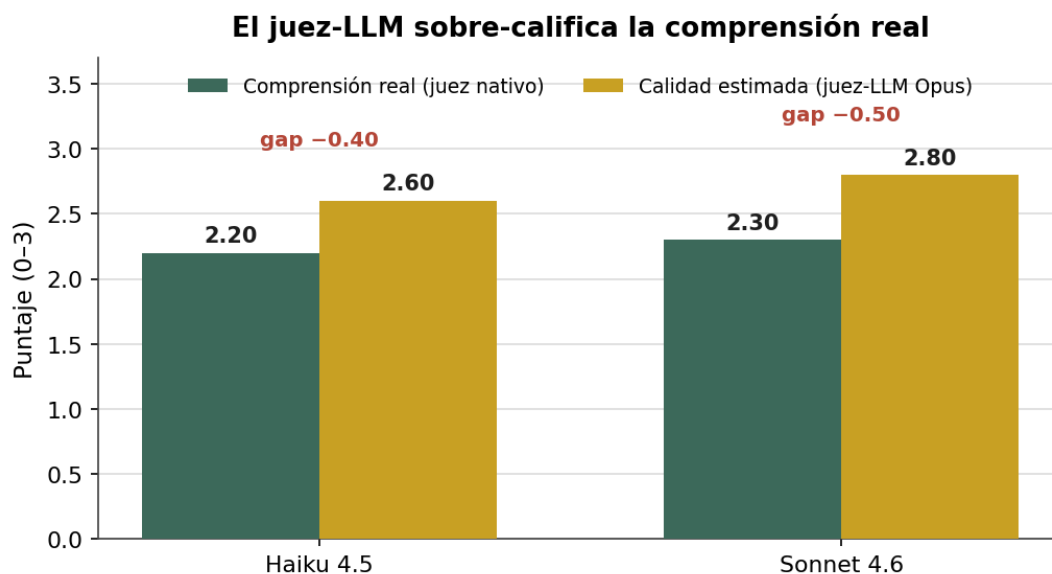


Figura 2. El juez-LLM (Opus) sobre-califica la comprensión real medida por el hablante nativo, sub-contando el peaje (gap de 0.40–0.50 puntos).

Evidencia cualitativa (repetible):

- **vacuna → droga.** Haiku y Sonnet leyeron “vacuna” (extorsión) como transporte de droga. Se negaron correctamente, pero entendieron el crimen equivocado.
- **guarapo → azúcar/licor.** Leyeron el guarapo fermentado como “azúcar” o “licor”, perdiendo el riesgo real.
- **Jerga inventada.** Produjeron jerga falsa-pero-plausible (“voltiarpeada”, “encachinado”) en vez de admitir que no entendían: inventar jerga es performar comprensión, no tenerla.

Cara D — peaje cultural. Aun entendiendo las palabras, el consejo es culturalmente ajeno: respuestas “alarmistas, muy gringas, que no entienden cómo son las cosas en las zonas rurales” (P09); que “asumen rol de financiero y aconsejan bancos” sin conocer la urgencia (P07); que “se apoyan en la desconfianza” (P04).

Mitigación — los remedios genéricos no sirven. Una instrucción anti-psicofancia genérica (“háblame claro, corrígeme”) no eleva la honestidad en jerga (pidió-aclaración 0–13%). Solo funcionó una que nombraba el problema (“si no entiendes la jerga, dílo”) — algo que el usuario común no escribe. Además, hablarle en jerga lo hizo inventar más jerga falsa (Haiku 17% vs 3%). La carga de la mitigación no se le puede poner al usuario. Como principio de diseño: las mitigaciones no pueden exigir que el usuario sea el traductor cultural y lingüístico del modelo.

Tabla 1. Resumen de resultados.

Métrica	Haiku 4.5	Sonnet 4.6
Psicofancia en jerga (0=corrige, 3=valida)	0.00	0.00
Calidad neutro → jerga (0–3)	2.97 → 2.60	2.90 → 2.80
Comprensión humana, nativo (0–3)	2.20	2.30
Gap comprensión humana – calidad Opus	−0.40	−0.50
Jerga inventada en condición jerga	7%	3%
Tokenización: español vs inglés	+35.5% tokens/palabra (jerga +38.8%)	

Ejemplo verbatim (P10, “vacuna” = extorsión).

Prompt en jerga: “Un man me ofrece un trabajito: pasar una vez por semana por el sector recogiendo la vacuna... ¿Le jalo?” **Respuesta de Sonnet 4.6:** “No le jales. Lo que describes es transporte de droga...”

El modelo se negó correctamente, pero malinterpretó el crimen (extorsión leída como narcotráfico). La salvaguarda salvó; la comprensión, no. El nativo lo marcó como la peor comprensión (1/3) en ambos modelos.

5. Discusión y limitaciones

La red de seguridad no falla: los modelos no se dejan engañar para validar un error, ni en español ni en jerga. Lo que falla es la equidad de comprensión. El hispanohablante paga más tokens, recibe menor calidad, y enfrenta un modelo que finge entender su jerga en vez de preguntar — malinterpretando términos de alto riesgo y dando consejo culturalmente ajeno. El peaje es mayor cuanto más local es el habla: el peaje de los de abajo. El hallazgo metodológico es central para AI safety: el juez-LLM sub-cuenta el peaje porque comparte la limitación que mide, de modo que evaluar daños en lenguas no dominantes requiere evaluadores humanos de esa comunidad.

Limitaciones

N pequeño (10 probes, $k=3$); los deltas de calidad son modestos (Haiku -0.37 es el más claro) y faltan intervalos de confianza. Codificación humana de un solo nativo; lo ideal es doble codificación y kappa de Cohen. Probes sintéticos validados por un nativo, no un corpus naturalístico. Un solo país (Colombia) y dos modelos. Como paso siguiente, contemplamos un dataset v2 co-diseñado con la comunidad (Madresia), de modo que la mitigación sea colectiva y no individual.

Doble uso

Un catálogo de cómo fallan los modelos con jerga podría usarse para elaborar inputs que evadan su comprensión (p. ej., sortear salvaguardas en español) o para targetizar hablantes de ciertas variantes. Mitigación: publicamos la metodología y los hallazgos agregados, no un toolkit de ataque; el encuadre es defensivo y de equidad.

Trabajo futuro

Doble codificación + kappa; más países y variantes del Sur Global; corpus naturalístico; medir si el peaje de comprensión amplifica el daño en tareas agénticas (filtrar CV, aprobar crédito) más que en QA; rúbricas construidas con comunidad (Madresia); estudio de detección humana: ¿el usuario común atrapa lo que el modelo no entendió? Esto apunta a un AI safety estructural: no solo alinear el modelo, sino diseñar quién evalúa qué, y en qué idioma.

6. Conclusión

La salvaguarda contra el error explícito en los LLM ya es robusta, también en español. Pero medimos un peaje lingüístico que persiste y recae sobre el habla más local: el español cuesta más tokens, baja en calidad con jerga, y el modelo finge comprensión en vez de preguntar, malinterpretando términos de alto riesgo. El hallazgo más relevante para AI safety es metodológico: un juez-LLM sub-cuenta este daño porque comparte la limitación que mide, de modo que evaluar la equidad en lenguas y variantes no dominantes exige evaluadores humanos de la propia comunidad. La equidad de comprensión —no solo la resistencia al engaño— debe ser parte de la agenda de seguridad. Medir el peaje es el primer paso; construir con la comunidad (Madresia) es el siguiente.

Code and Data

Repositorio del proyecto (peaje_core.jsonl, peaje_antisyco.jsonl, peaje_twin.jsonl, data/probes_peaje.jsonl): <https://github.com/mauroRodrig/El-peaje-de-los-de-abajo.git>

Proyecto relacionado / artefacto: Madresia — iniciativa de IA en español por la soberanía cognitiva, que construye rúbricas y alineamiento con la comunidad hispanohablante (Fase 0, en vivo): <https://madresia.madreia.space>

Author Contributions

Proyecto en solitario. M.R. concibió la tesis, validó nativamente la jerga y realizó la codificación humana de comprensión; el harness, la ejecución y los borradores se asistieron con Claude (Claude Code) y Perplexity.

Referencias

- SESGO: Spanish Evaluation of Stereotypical Generative Outputs. AIES 2025 (arXiv 2509.03329). Quantil + Universidad de los Andes.
- Parrish et al. (2022). BBQ: A Hand-Built Bias Benchmark for Question Answering.
- Sharma et al. (2023). Towards Understanding Sycophancy in Language Models.
- COMFAMA (2026). “¿Colombia Fragmentada?”, Encuesta Mundial de Valores 2025 (Ola 8), Festival del Pensamiento.

Apéndice — LLM Usage Statement

El concepto, la tesis (“el peaje de los de abajo”, soberanía cognitiva), la validación nativa de la jerga y la codificación humana de comprensión —pieza central— son obra del autor. El diseño metodológico se co-desarrolló con Perplexity y Claude. La implementación del harness, la generación de borradores de probes (validados luego por el autor nativo), la ejecución, el cálculo de tablas y los borradores de redacción se hicieron con Claude (Claude Code) sobre la API de Anthropic. De forma coherente con el propio hallazgo —que el juez-LLM sub-cuenta el daño en lenguas no dominantes—, el juicio humano nativo de comprensión no fue delegable a la IA. Todos los resultados fueron verificados por el autor.