DOCUMENTO TÉCNICO · EQUIPO LEXIMEX

IA Legal Verificable

Metodología, arquitectura anti-alucinación y benchmark para asistencia jurídica sobre derecho mexicano

Versión 1.0 · Julio de 2026

Cobertura evaluada: legislación federal + 32 entidades federativas

Benchmark anti-alucinación: n = 108 preguntas · 10 áreas del derecho

Estadística: intervalos de confianza de Wilson (95%)

Tasa de alucinación medida: 5,6% (IC 95%: 2,6%–11,6%) · 0 en 87 preguntas de uso real

LEXIMEX Inteligencia artificial legal para abogados de México · leximex.com

RESUMEN

Los sistemas de IA generativa aplicados al derecho sufren de *alucinación*: inventan artículos, tesis o criterios que no existen. Estudios independientes (Stanford, 2024–2025) documentan tasas de 17% a 33% en herramientas comerciales líderes, incluso con recuperación aumentada por documentos (RAG). LEXIMEX es una plataforma de asistencia jurídica sobre derecho mexicano construida alrededor de un principio inverso: **toda afirmación normativa se verifica contra una base de datos legal propia antes de mostrarse, y cuando no hay fundamento, el sistema lo declara en lugar de improvisar.** Este documento describe la arquitectura de verificación de cinco capas, la metodología de evaluación y los resultados de un benchmark de 108 preguntas —incluidas 21 preguntas-trampa adversariales— sobre 10 áreas del derecho. La tasa de alucinación medida fue de **5,6%** (6/108; IC 95% de Wilson: 2,6%–11,6%), concentrada en su totalidad en las preguntas-trampa; en las 87 preguntas de uso realista no se detectó ninguna alucinación (cota superior del IC 95%: 4,2%). No afirmamos una tasa de cero: publicamos el número real, su intervalo de confianza y sus límites.

1. Introducción: el problema de la alucinación en IA legal

Un modelo de lenguaje generativo optimiza la *plausibilidad* del texto, no su *veracidad*. Cuando se le pregunta por el artículo 999 de la Ley Federal del Trabajo —que no existe— tenderá a redactar una respuesta convincente antes que a admitir que la norma no existe. En cualquier otro dominio esto es un inconveniente; en el ejercicio del derecho es una fuente de responsabilidad profesional. Un escrito fundado en un artículo inexistente o en una tesis inventada puede costar un plazo, una audiencia o la credibilidad del abogado ante un juez.

El estudio de la Universidad de Stanford *«Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools»* (Magesh et al., 2024–2025) evaluó herramientas comerciales de investigación jurídica con arquitectura RAG y midió tasas de alucinación de entre 17% y 33%. La conclusión relevante es que **añadir recuperación de documentos reduce el problema, pero no lo elimina**: el modelo sigue teniendo la última palabra sobre qué escribe, y esa última palabra puede desviarse del documento recuperado.

LEXIMEX parte de una premisa distinta: el modelo redacta, pero **no es la autoridad final sobre qué es verdad**. Cada cita a un artículo o tesis atraviesa una tubería de verificación determinista contra la base de datos legal antes de llegar al usuario. Este documento explica cómo funciona esa tubería y cuánto mueve la aguja, con los números medidos y sus intervalos de confianza.

2. Trabajo relacionado

La literatura sobre confiabilidad de IA legal converge en tres hallazgos. Primero, la alucinación persiste incluso en sistemas RAG de nivel profesional (Stanford, 2024–2025). Segundo, los modelos son sistemáticamente *sobreconfiados*: expresan certeza alta en respuestas incorrectas, por lo que la confianza autodeclarada del modelo no es un indicador fiable. Tercero, la verificación efectiva requiere una fuente de verdad externa y determinista —una base de datos consultable— contra la cual contrastar cada afirmación, en lugar de confiar en que el modelo se autocorrija.

La arquitectura de LEXIMEX operacionaliza estos tres hallazgos: (i) asume que el modelo alucinará y diseña para atraparlo; (ii) reemplaza la confianza autodeclarada por una confianza *computada* a partir de señales objetivas; y (iii) usa la base de datos legal —no el modelo— como árbitro final de existencia de cada norma citada.

3. Arquitectura: cinco capas de verificación

Toda respuesta generada por un agente de LEXIMEX atraviesa una tubería de verificación (`verify_response`) antes de mostrarse. Las capas no se sustituyen: se acumulan.

3.1 Por qué existe cada capa

Capa	Qué hace	Qué falla si no existiera
1 · Verificación de citas	Extrae cada artículo/tesis citado y comprueba su existencia real en la base de datos legal.	El sistema citaría el «artículo 999 LFT» sin que nadie confirme que existe.
2 · Confianza computada	Calcula una confianza objetiva a partir de la similitud RAG y la tasa de citas verificadas, no de lo que el modelo *dice* sentir.	Se mostraría la certeza del modelo, que es alta incluso cuando se equivoca.
3 · Filtro anti-alucinación	Bloquea o degrada respuestas cuya confianza computada o proporción de citas verificadas caen por debajo del umbral.	Una respuesta mal fundada llegaría al usuario con apariencia de válida.
4 · Contraste con el contexto RAG	Verifica que cada cita esté efectivamente presente en los fragmentos recuperados, no solo que «exista en algún lugar».	El sistema podría citar de memoria un artículo real pero fuera de contexto para la pregunta.
5 · Señales de alucinación	Detecta patrones de riesgo (citas sin anclaje, incoherencias) y penaliza la confianza en consecuencia.	Fallos sutiles pasarían sin penalización aunque las citas «existieran».

El resultado no es un texto pulido con un número inventado de confianza, sino un texto acompañado de un **sello auditable**: «X/Y artículos verificados en contexto», nivel de confianza (ALTA / MEDIA / BAJA) y advertencias explícitas cuando algo no cuadra. Cuando la evidencia no alcanza, el sistema prefiere decir *«no encontré fundamento»* antes que rellenar el hueco.

3.2 El mismo principio en el Simulador de Juicios

La verificación no es exclusiva del chat. El Simulador de Juicios de LEXIMEX —que analiza la viabilidad de un caso desde tres roles (juez, contraparte y estratega)— aplica la misma tubería *por sección*: cada apartado del análisis lleva su propio recuento de artículos verificados. Un análisis de estrategia procesal con «26/27 citas verificadas» es literalmente auditable artículo por artículo.

4. Metodología de evaluación

4.1 Banco de preguntas

El benchmark v1.0 consta de **108 preguntas** distribuidas en 10 áreas del derecho y seis categorías diseñadas para estresar distintos modos de fallo. La clave del diseño son las **preguntas-trampa** (TRAP): preguntas que citan artículos inexistentes con apariencia verosímil (p. ej. «según el artículo 999 de la LFT…») para medir si el sistema se deja arrastrar o corrige.

Categoría	Propósito	Preguntas
NORMAL	Consultas jurídicas reales, respuesta con fundamento existente	42
TRAP (adversarial)	Cita un artículo/tesis falso para inducir alucinación	21
EDGE	Casos límite, cruces de competencia, supuestos poco frecuentes	15
AMBIGUOUS	Preguntas ambiguas que requieren pedir precisión	10
HONESTY	Preguntas cuya respuesta correcta es «no hay fundamento / no lo sé»	10
CROSS_REF	Requieren cruzar más de una norma o jurisdicción	10
Total		**108**

4.2 Rúbrica y calificación

Cada respuesta se calificó de 0 a 100 combinando: (a) corrección jurídica del fundamento, (b) que toda cita citada exista y esté en contexto, (c) honestidad —penalizando la invención y premiando el «no sé» cuando corresponde— y (d) detección de la trampa en las preguntas TRAP. Una respuesta que cae en la trampa (cita el artículo falso como real) se marca como **alucinación**, con independencia de lo bien redactada que esté.

4.3 Tratamiento estadístico

Reportar «5,6%» sin más sería estadísticamente incompleto: con 108 preguntas, ese punto tiene incertidumbre. Para cada proporción calculamos el **intervalo de confianza de Wilson al 95%**, apropiado para proporciones con muestras moderadas y para el caso de cero eventos (donde la fórmula normal se rompe). Todos los resultados se presentan como *estimación puntual + intervalo*, nunca como número aislado.

4.4 Re-validación continua

El benchmark v1.0 es una fotografía. Para vigilar regresiones tras cambios de modelo o de recuperación, LEXIMEX ejecuta un **arnés de evaluación semanal** (`run_eval.py`) sobre un conjunto *golden* de preguntas con respuesta esperada conocida, midiendo tasa de aprobación, *recall* de citas y proporción de citas verificadas. Es el equivalente a un análisis de sangre periódico: no sustituye al estudio completo, pero avisa temprano si algo se degrada.

5. Resultados

5.1 Benchmark principal (26 de febrero de 2026)

94,4
Score promedio /100

93,5%
Aprobación (101/108)

5,6%
Alucinación global (6/108)

0,88
Similitud RAG media

Métrica	Resultado	IC 95% (Wilson)
Tasa de alucinación — global	**5,6% (6/108)**	2,6% – 11,6%
Tasa de alucinación — **uso realista** (87 preg. no-trampa)	**0,0% (0/87)**	0,0% – 4,2%
Tasa de alucinación — preguntas-trampa adversariales	**28,6% (6/21)**	13,8% – 50,0%
Tasa de aprobación (score ≥ umbral)	93,5% (101/108)	87,2% – 96,8%

El hallazgo central es honesto, no perfecto. En las 87 preguntas que reflejan uso real —consultas normales, casos límite, cruces de normas y preguntas de honestidad— el sistema no inventó ninguna norma (0/87). Las 6 alucinaciones ocurrieron todas en las 21 preguntas-trampa: ataques deliberados que citan artículos inexistentes. Es decir, cuando el sistema falla, falla bajo agresión adversarial explícita, no en el trabajo cotidiano del abogado. Y aun bajo ese ataque, la tasa global de 5,6% es entre 3 y 6 veces menor que la de las herramientas comerciales medidas por Stanford.

5.2 Desglose por categoría

Categoría	Preg.	Aprobadas	Alucinaciones	Score
NORMAL	42	42	**0**	96,5
TRAP	21	14	**6**	81,8
AMBIGUOUS	10	10	**0**	100,0
EDGE	15	15	**0**	97,1
HONESTY	10	10	**0**	97,0
CROSS_REF	10	10	**0**	100,0

5.3 Referencia con la literatura (Stanford, 2025)

Herramienta	Tasa de alucinación	Fuente
Ask Practical Law AI	**33,0%**	Stanford 2025
Westlaw AI-Assisted Research	**25,0%**	Stanford 2025
Lexis+ AI	**17,0%**	Stanford 2025
GPT-4 sin RAG	**~30–40%**	Estimación industria
LEXIMEX (global, con trampas)	**5,6%**	Este benchmark
LEXIMEX (uso realista)	**0,0%**	Este benchmark

Nota de comparabilidad: los benchmarks de Stanford y de LEXIMEX usan bancos de preguntas y jurisdicciones distintos (derecho estadounidense vs. derecho mexicano), por lo que la comparación es *orientativa*, no un contraste controlado. Se incluye para situar el orden de magnitud, no para afirmar superioridad estadística directa.

5.4 Re-validación tras migración de modelo

Tras la migración del modelo de razonamiento para consultas complejas, el arnés semanal sobre el conjunto *golden* registró (medición del 5 de julio de 2026): tasa de aprobación 90%, *recall* de citas 0,87 y **proporción de citas verificadas 1,00** —es decir, el 100% de las citas mostradas existían y estaban en contexto—. La verificación se mantuvo estable a través del cambio de modelo, que es precisamente lo que el arnés está diseñado para garantizar.

6. Limitaciones declaradas y trabajo futuro

Publicar un benchmark obliga a publicar también sus límites. Los nuestros:

- **La tasa no es cero y no la prometemos así.** Bajo ataque adversarial explícito, 6 de 108 respuestas cayeron en la trampa. Cualquier proveedor que afirme «0% de alucinación» está midiendo mal o comunicando mal.
- **Tamaño muestral moderado.** Con n = 108, el IC 95% de la tasa global va de 2,6% a 11,6%. Ampliar el banco a n ≥ 250 estrecharía el intervalo; es trabajo comprometido.
- **Calificación asistida.** La rúbrica se aplicó de forma consistente, pero la validación humana independiente con acuerdo entre evaluadores (kappa de Cohen) está pendiente y es el siguiente hito de rigor.
- **Comparación con Stanford orientativa.** Distinta jurisdicción y distinto banco de preguntas; sirve para el orden de magnitud, no como prueba estadística directa.
- **Cobertura por diseño.** LEXIMEX indexa la legislación federal y los códigos y leyes principales de las 32 entidades (435 ordenamientos, 162.571 fragmentos, 31.857 tesis de la SCJN a la fecha de este documento). No pretende contener cada reglamento municipal; cuando una norma no está indexada, el sistema lo declara en vez de inventarla.

El trabajo futuro incluye: ampliación del banco a n ≥ 250, validación humana con kappa, re-medición trimestral con el mismo arnés y pre-registro público de la metodología antes de cada nueva medición.

7. Conclusión

La pregunta correcta para una IA legal no es «¿alucina?» —todas lo hacen en algún grado— sino «¿cuánto, bajo qué condiciones, y lo publica?». LEXIMEX responde con un número medido (5,6% global; 0/87 en uso realista), su intervalo de confianza y sus límites. La diferencia de diseño es simple de enunciar: **el modelo redacta, pero la base de datos legal decide qué es verdad**, y cuando no hay fundamento el sistema lo dice. Esa disciplina —verificar antes de afirmar, y declarar la incertidumbre en lugar de esconderla— es lo que separa una herramienta en la que un abogado puede fundar un escrito de una que solo suena convincente.

Reproducibilidad. El banco de preguntas, el arnés de evaluación y los resultados por categoría están disponibles para revisión bajo solicitud a través de leximex.com. La postura del equipo es que quien afirme una cifra de precisión debe poder mostrar el benchmark que la sostiene.

8. Referencias

1. Magesh, V., Barrett, F., Miller, T., et al. (2024–2025). *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools.* Stanford University — RegLab & Institute for Human-Centered AI.
2. Wilson, E. B. (1927). *Probable inference, the law of succession, and statistical inference.* Journal of the American Statistical Association, 22(158), 209–212.
3. Equipo LEXIMEX (2026). *Benchmark anti-alucinación v1.0 — protocolo, banco de preguntas y resultados.* Documentación interna, leximex.com.