ABOGADOIA · **ABOGADOIA.CL** · PLATAFORMA DE IA PARA EL DERECHO CHILENO

IA legal verificable para el derecho chileno: arquitectura antialucinación de cinco capas y su evaluación empírica sobre 115 consultas

Documento técnico-científico. Describe el problema de la **alucinación en asistentes jurídicos de IA** y su costo profesional, presenta una arquitectura de generación aumentada por recuperación (RAG) sobre un corpus verificado de legislación y jurisprudencia chilena con **verificación post-generación de cada cita**, y reporta los resultados de un benchmark interno de 115 consultas en 12 áreas del derecho: **0 alucinaciones detectadas**, lo que acota la tasa estimada a **<3,5% con 95% de confianza**. Se declaran las limitaciones del método y el trabajo futuro comprometido.

DOCUMENTO TÉCNICO · v1.0 · No revela componentes propietarios del sistema

Elaborado por
Equipo AbogadoIA

Contacto
abogadoia.cl · contacto@abogadoia.cl

Fecha
Julio 2026

RESUMEN

Los modelos de lenguaje de gran escala inventan normas, artículos y jurisprudencia con fluidez convincente. El estudio de referencia del campo (Stanford RegLab / Yale, 2024) midió tasas de alucinación de ~17% en Lexis+ AI y ~33% en Westlaw AI-Assisted Research, herramientas comerciales líderes que se presentaban como confiables [1]. Este documento describe el sistema antialucinación de **AbogadoIA**, una plataforma de IA especializada en derecho chileno: una arquitectura de **cinco capas** que combina (i) un contrato de comportamiento impuesto al modelo, (ii) generación fundada exclusivamente en un corpus legal verificado de 53 leyes chilenas (19.024 fragmentos a nivel de artículo) y 6.699 sentencias, recuperado mediante búsqueda híbrida léxico-semántica con fusión de rankings, (iii) verificación post-generación de cada cita contra el corpus, (iv) un puntaje de confianza compuesto con señales de alucinación y (v) un aviso de transparencia calibrado a ese puntaje. En un benchmark interno de **115 consultas** sobre los 12 agentes especializados de la plataforma (6 categorías, incluidas 22 preguntas trampa con normas inexistentes y 12 de honestidad), el sistema registró **0 alucinaciones detectadas**, 100% de detección de trampas (22/22), 100% de honestidad (34/34) y un puntaje promedio de 97,7/100. Cero eventos en 115 observaciones no permiten afirmar una tasa nula: la cota superior del intervalo de Wilson al 95% es 3,2%, por lo que el resultado se comunica como **tasa de alucinación estimada <3,5% (IC 95%)**. Una re-validación de julio de 2026, tras una migración de modelo, confirmó recall de artículos y de leyes de 100% sobre las mismas 115 preguntas y, tras una corrección de reglas de premisas falsas, 22/22 en trampas y 12/12 en honestidad. Se declaran las limitaciones (n=115, evaluación con juez automatizado, banco de preguntas interno) y los siguientes hitos del programa de evaluación.

1 Introducción: el problema de la alucinación en IA legal

La alucinación —la generación de afirmaciones falsas presentadas con seguridad— no es un defecto accidental de los modelos de lenguaje: es una propiedad estadística de su modo de operación. En el dominio jurídico su forma más peligrosa es concreta y reconocible: **citar un artículo que no existe, atribuir a una ley un contenido que no tiene, o inventar jurisprudencia con roles y carátulas plausibles**. A diferencia de otros dominios, aquí el error no se diluye: se incorpora a un escrito, se presenta ante un tribunal y compromete la responsabilidad profesional del abogado que lo firma.

Las consecuencias ya no son hipotéticas. Desde *Mata v. Avianca* (S.D.N.Y., 2023, sanción por seis precedentes ficticios generados por un chatbot de consumo), los tribunales de múltiples jurisdicciones han sancionado presentaciones con citas fabricadas por IA, y Chile no es la excepción: en 2026 se registran sanciones de tribunales chilenos a abogados por invocar jurisprudencia inexistente generada con IA. El riesgo, además, es asimétrico: el profesional no puede auditar a simple vista una cita bien formateada, y la revisión manual de cada referencia anula gran parte del ahorro de tiempo que motiva el uso de la herramienta.

La posición del equipo de AbogadoIA es que este problema **no se resuelve pidiéndole al modelo que no alucine, sino imponiendo la confiabilidad desde la arquitectura**: restringir lo que el modelo puede afirmar a un corpus verificado, comprobar por código cada cita emitida antes de que llegue al usuario, y comunicar con honestidad el nivel de confianza de cada respuesta, incluida la abstención cuando el corpus no contiene la respuesta. Este documento describe esa arquitectura (§3), la metodología con que se midió (§4), sus resultados (§5) y sus limitaciones (§6).

2 Trabajo relacionado

El punto de referencia empírico del campo es el estudio de Stanford RegLab y Yale (*Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools*; Magesh, Surani, Dahl, Suzgun, Manning y Ho, 2024; publicado en *Journal of Empirical Legal Studies*, 2025) [1]. Sobre más de 200 consultas preregistradas y con calificación por expertos, midió que las principales plataformas comerciales de investigación legal con IA alucinan con frecuencia significativa: **~17% en Lexis+ AI y ~33% en Westlaw AI-Assisted Research**, según sus mediciones y definiciones. Su hallazgo central es doble: el RAG reduce la alucinación respecto de los chatbots de propósito general (que los mismos autores habían medido entre 58% y 82% en consultas jurídicas [2]), pero **no la elimina**, y las afirmaciones comerciales de estar "libre de alucinaciones" no resistieron la medición independiente.

De ese estudio el equipo adopta dos lecciones de diseño: (a) la evaluación debe incluir **preguntas trampa** con premisas falsas y normas inexistentes, porque es allí donde los sistemas RAG fallan de forma característica (recuperan algo parecido y lo afirman); y (b) ningún proveedor debe declarar tasas nulas — la nuestra se reporta como una cota estadística, no como un cero (§5). Advertimos desde ya que **las cifras de Stanford no son directamente comparables con las de este documento**: aquellas provienen de consultas de investigación jurídica estadounidense calificadas por expertos humanos; las nuestras, de un banco interno sobre derecho chileno con evaluación automatizada (§6). La comparación es de orden de magnitud y de metodología, no una equivalencia.

3 Arquitectura del sistema: cinco capas antialucinación

AbogadoIA opera 12 agentes especializados, uno por área del derecho chileno (civil, penal, laboral, tributario, familia, administrativo, comercial, constitucional, inmobiliario, consumidor, minero y propiedad intelectual). Todos comparten el mismo pipeline: un modelo de lenguaje de gran escala con razonamiento extendido, encapsulado dentro de cinco capas donde **ninguna afirmación con cita llega al usuario sin haber sido contrastada contra el corpus**.

1 · Contrato de comportamiento	Instrucciones de fundamentación que envuelven cada consulta: el agente solo puede afirmar lo sostenido por el material recuperado, debe declarar cuando el corpus no contiene la respuesta y debe corregir explícitamente las premisas falsas del usuario (normas o artículos inexistentes).
▼	
2 · Generación fundada (RAG híbrido sobre corpus verificado)	Corpus de 53 leyes chilenas segmentadas a nivel de artículo (19.024 fragmentos) + 6.699 sentencias. Recuperación híbrida: búsqueda léxica (BM25) + semántica (embeddings vectoriales) fusionadas por ranking recíproco (RRF), con expansión de consulta y filtro por área del derecho. Embeddings calculados localmente: la consulta no sale del servidor para vectorizarse.
▼	
3 · Verificación post-generación de citas	Cada artículo y ley citados en la respuesta se contrastan contra el corpus por código, no por otro modelo: existencia de la norma, existencia del artículo y consistencia con el texto recuperado.
▼	
4 · Puntaje de confianza compuesto	Combina el resultado de la verificación de citas, la calidad de la recuperación (similitud del contexto) y señales lingüísticas de alucinación detectadas en la respuesta. La verificación determinista domina sobre la fluidez del modelo.
▼	
5 · Transparencia calibrada	La respuesta se entrega con su nivel de confianza (ALTA/MEDIA/BAJA) y un aviso proporcional: una respuesta de confianza BAJA se marca visiblemente como tal antes del texto, incluso en modo streaming.

3.1 Por qué cada capa existe

Capa 1 — el modelo no decide su propio estándar de verdad. El contrato de comportamiento convierte la abstención en el fallo seguro: ante una consulta cuya respuesta no está en el corpus, la conducta esperada —y medida en el benchmark (categoría HONESTY, §4)— es declararlo, no improvisar. Una regla específica obliga a corregir premisas falsas: si el usuario pregunta por la "Ley de Protección al Teletrabajador Digital" y esa ley no existe, el agente debe decirlo y reconducir a la norma real, no seguir la corriente.

Capa 2 — la mayoría de las alucinaciones nacen como fallas de recuperación. La búsqueda puramente semántica rendía una similitud promedio de 0,76 en terminología legal chilena: los embeddings pierden precisión justamente en lo que el derecho exige exacto (números de artículo, números de ley, términos técnicos). El canal léxico BM25 captura esas coincidencias exactas; el semántico, los sinónimos y el parafraseo. La fusión por ranking recíproco de ambos canales elevó la similitud promedio del contexto recuperado a 0,893 en el benchmark (§5). Los agentes filtran además el corpus por área del derecho, lo que reduce la recuperación de material plausible pero impertinente.

Capa 3 — el verificador no es un modelo. Es la barrera diferencial: un componente determinista, auditable línea a línea, recalcula cada cita contra la base de datos. Un modelo de lenguaje puede redactar una cita falsa convincente, pero no puede "convencer" a una consulta SQL de que un artículo existe. Lo que el verificador no confirma, degrada el puntaje de confianza de la respuesta completa.

Capas 4 y 5 — medir la confianza y decirla. El puntaje compuesto pondera verificación, calidad de recuperación y señales de alucinación; su calibración es deliberadamente conservadora (prefiere degradar una respuesta correcta antes que sobrevalorar una dudosa). El aviso final no es un descargo legal genérico: es proporcional al puntaje, y en confianza BAJA antecede visiblemente al texto de la respuesta.

3.2 El mismo principio en el Simulador de Juicios

La funcionalidad más reciente de la plataforma somete la arquitectura a una prueba más exigente. El Simulador de Juicios genera, **en paralelo, tres análisis independientes** de un caso: la estrategia propia, la posición esperable de la contraparte y la perspectiva del tribunal. Cada análisis se verifica **por separado, cita por cita, contra el corpus** (capa 3 aplicada por sección), y la confianza global de la simulación se define como **la peor de las tres secciones**: si el análisis de la contraparte contiene una cita no verificable, toda la simulación se degrada, aunque las otras dos secciones sean impecables. La regla de diseño es la misma de todo el sistema: **la verificación manda sobre el modelo** — ningún componente generativo puede elevar la confianza que la verificación determinista rebajó.

4 Metodología de evaluación

Benchmark interno ejecutado el 27 de febrero de 2026 contra el pipeline completo en su configuración de producción: **115 consultas**, los **12 agentes**, duración total 46,3 minutos. Diseño inspirado en el estudio de Stanford RegLab [1], adaptado al derecho chileno.

4.1 Banco de preguntas

Las 115 preguntas se distribuyen entre los 12 agentes (9–11 por área) y en **seis categorías**, diseñadas para medir modos de fallo distintos:

CATEGORÍA	N	QUÉ MIDE
NORMAL	42	Consultas legales típicas con respuesta verificable en el corpus: exactitud de artículos, leyes y contenido.
TRAP	22	Preguntas trampa: normas o artículos **inexistentes**, premisas falsas embebidas. El sistema debe detectar y corregir, no seguir la corriente.
EDGE	15	Casos límite: materias en la frontera del corpus o entre áreas del derecho.
HONESTY	12	Consultas cuya respuesta no está en el corpus: la conducta correcta es la abstención declarada.
AMBIGUOUS	12	Consultas ambiguas que admiten más de una lectura: el sistema debe explicitar la ambigüedad.
CROSS_REF	12	Consultas que exigen cruzar dos o más cuerpos legales correctamente.

4.2 Rúbrica y calificación

Cada respuesta se puntúa de 0 a 100 combinando: (i) **corrección de citas** — todo artículo citado se contrasta contra el corpus, exigiendo que la ley correspondiente aparezca en la vecindad inmediata de la cita; (ii) **cobertura** de los artículos y leyes esperados para la pregunta (recall de artículos y de leyes); (iii) **detección de trampa** (binaria, en TRAP); (iv) **honestidad** (binaria: abstención o advertencia explícita donde corresponde); y (v) **calibración de confianza** — que el nivel ALTA/MEDIA/BAJA declarado sea coherente con el resultado de la verificación. Se considera **alucinación** la afirmación de una norma o artículo inexistente, la atribución de contenido falso a una norma real, o la aceptación de la premisa falsa de una pregunta trampa. La calificación es automatizada (contraste determinista contra la base de datos más reglas de evaluación) con revisión posterior de los casos limítrofes; sus limitaciones se declaran en §6.

4.3 Tratamiento estadístico

Con 0 eventos observados en n=115, la estimación puntual de la tasa de alucinación es 0%, pero ese número **no es publicable como claim**: la incertidumbre de muestreo se cuantifica con el intervalo de Wilson, recomendado para proporciones extremas con n moderado. Para 0/115 al 95% de confianza, la cota superior es **3,2%**. Por eso todo material público de AbogadoIA comunica el resultado como **"tasa de alucinación estimada <3,5% (IC 95%; 0 detectadas en 115 preguntas)"** — nunca como "0%" ni "cero alucinaciones".

> **Por qué no prometemos "cero alucinaciones".** Es estadísticamente indefendible: incluso 0 fallos en cientos de pruebas deja una cota superior no nula. Y es la lección central de Stanford [1]: las herramientas que se vendían como "hallucination-free" midieron 17% y 33%. Nuestro compromiso es el contrario — publicar la tasa con su intervalo, sus condiciones de medición y sus limitaciones, y re-medir con el mismo arnés ante cada cambio de modelo o de prompt.

4.4 Re-validación continua (arnés de evaluación)

El banco de 115 preguntas se convirtió en un **arnés de evaluación permanente** que corre contra el pipeline real y directo (sin caché ni capas de respuesta rápida, para no enmascarar cambios). Ante cualquier modificación de modelo o de instrucciones, se ejecuta una

comparación pareada antes/después sobre las mismas preguntas; **una regresión en tasa de alucinación o en detección de trampas bloquea el despliegue**. La primera aplicación fue la migración de modelo de julio de 2026 (§5.2).

5 Resultados

5.1 Benchmark principal (27 de febrero de 2026)

0/115	22/22	34/34	97,7
ALUCINACIONES DETECTADAS → TASA ESTIMADA <3,5% (IC 95%)	DETECCIÓN DE TRAMPAS (NORMAS INEXISTENTES)	HONESTIDAD (ABSTENCIÓN/ADVERTENCIA DONDE CORRESPONDE)	PUNTAJE PROMEDIO /100 · 115/115 APROBADAS

MÉTRICA	RESULTADO
Preguntas evaluadas / aprobadas	**115 / 115 (100%)**
Alucinaciones detectadas	**0** → cota superior Wilson 95%: **3,2%**
Detección de trampas (TRAP)	22/22 (100%)
Honestidad (HONESTY + casos aplicables)	34/34 (100%)
Puntaje promedio	97,7/100
Similitud promedio del contexto recuperado (RAG)	0,893
Tiempo promedio de respuesta	17,6 s

Resultados por categoría

CATEGORÍA	N	APROBADAS	ALUCINACIONES	PUNTAJE
NORMAL	42	42	0	95,4
TRAP	22	22	0	99,5
AMBIGUOUS	12	12	0	100,0
EDGE	15	15	0	98,0
HONESTY	12	12	0	99,7
CROSS_REF	12	12	0	97,8

Los 12 agentes aprobaron el 100% de sus preguntas, con puntajes entre 95,7 (constitucional) y 99,5 (laboral). Reporte completo: `benchmark_20260227_022824`.

5.2 Re-validación tras migración de modelo (4 de julio de 2026)

En julio de 2026 la plataforma migró a una nueva generación del modelo de lenguaje subyacente. El arnés (§4.4) corrió las 115 preguntas contra el pipeline actualizado: **recall de artículos 100% y recall de leyes 100%**, similitud de recuperación 0,919 y **una única alucinación parcial** detectada — en una pregunta trampa, el agente citó la norma real aplicable pero omitió corregir explícitamente el nombre de la ley inventada por el usuario. La corrección no fue un parche puntual sino una regla general del contrato de comportamiento (capa 1): la obligación explícita de identificar y corregir premisas falsas. Tras esa regla, la re-ejecución de las categorías afectadas arrojó **TRAP 22/22 con 0 alucinaciones y HONESTY 12/12**, restituyendo el respaldo vigente del resultado 0/115 sobre el pipeline en producción.

CORRIDA	N	ALUCINACIÓN	TRAMPAS	HONESTIDAD	RECALL ART. / LEYES

Re-validación completa (pre-corrección)	115	1 parcial (0,9%)	21/22	94,1%	100% / 100%
Post-corrección — TRAP	22	**0**	**22/22**	100%	—
Post-corrección — HONESTY	12	**0**	—	**12/12**	—

El episodio ilustra el valor del arnés: la regresión se detectó y corrigió *antes* de que el cambio de modelo quedara sin medición, y la corrección se validó sobre las mismas preguntas.

5.3 Referencia con la literatura

SISTEMA	TASA DE ALUCINACIÓN	MEDICIÓN
Westlaw AI-Assisted Research	~33%	Stanford RegLab [1], expertos humanos, derecho de EE. UU.
Lexis+ AI	~17%	Stanford RegLab [1], expertos humanos, derecho de EE. UU.
Chatbots de propósito general	58–82%	Dahl et al. [2], consultas jurídicas, EE. UU.
AbogadoIA	**<3,5% (IC 95%)**	Benchmark interno, n=115, derecho chileno, juez automatizado

> **Honestidad metodológica.** Estas cifras **no son directamente comparables**: difieren el ordenamiento jurídico (EE. UU. vs. Chile), el tipo de consulta, el origen del banco de preguntas (preregistrado externo vs. interno) y sobre todo el calificador (expertos humanos vs. evaluación automatizada con revisión). La tabla se ofrece como referencia de orden de magnitud del problema que este sistema ataca, no como un ranking.

6 Limitaciones declaradas y trabajo futuro

1. **Tamaño muestral.** n=115 acota la tasa a <3,5% con 95% de confianza; no permite distinguir entre una tasa real de 0,5% y una de 3%. Hito comprometido: ampliar el banco a **n=500** para estrechar el intervalo.
2. **Calificador automatizado.** La evaluación es un contraste determinista contra el corpus más reglas de calificación, con revisión de casos limítrofes — no una calificación por abogados independientes. Falta la **validación humana de una muestra con reporte de acuerdo inter-anotador (kappa)**; está comprometida como siguiente hito del programa de evaluación.
3. **Banco de preguntas interno.** Fue diseñado por el mismo equipo que construyó el sistema. El plan incluye el **preregistro** de la próxima versión del banco (preguntas fijadas y publicadas antes de ejecutar la medición) para eliminar todo grado de libertad post hoc.
4. **No determinismo del modelo.** Los resultados corresponden a corridas puntuales; los modelos no son deterministas. Las comparaciones entre versiones se hacen pareadas sobre las mismas preguntas, nunca entre promedios de corridas distintas.
5. **Alcance del corpus.** El sistema responde sobre 53 leyes chilenas y 6.699 sentencias; el derecho chileno es más vasto. Fuera del corpus, la conducta correcta —y medida— es la abstención, pero un corpus mayor reduce la frecuencia de abstenciones. La ampliación del corpus es trabajo continuo y cada ampliación se re-mide con el arnés.
6. **Definición operativa de alucinación.** Se centra en citas e imputaciones normativas verificables contra el corpus. Errores jurídicos más sutiles (una interpretación doctrinaria discutible sin cita falsa) exceden la definición y requieren la validación humana del punto 2.

El sistema se diseña, en consecuencia, como **asistente del profesional, no como sustituto de su juicio**: toda respuesta entrega sus citas verificadas y su nivel de confianza precisamente para que el abogado pueda auditar en segundos lo que firma.

7 Conclusión

La alucinación en IA legal está medida, sancionada y sin resolver en los líderes del mercado. La respuesta de AbogadoIA no es una promesa de infalibilidad sino una arquitectura donde la confiabilidad es una propiedad estructural: un corpus legal chileno verificado como única fuente de verdad, recuperación híbrida léxico-semántica que preserva la exactitud que el derecho exige, verificación por código de cada cita antes de llegar al usuario, un puntaje de confianza donde la verificación determinista domina sobre el modelo, y transparencia calibrada — extendida sección por sección al Simulador de Juicios, donde la confianza global es la de la peor sección. La evidencia disponible (0 alucinaciones detectadas en 115 consultas, 100% de detección de trampas y de honestidad, re-validada tras la migración de modelo de julio de 2026) sostiene el claim público de una **tasa de alucinación estimada inferior a 3,5% con 95% de confianza** — con sus limitaciones declaradas y un programa de evaluación comprometido para estrecharlas: validación humana con kappa, banco preregistrado y n=500. En un mercado donde nadie publica sus tasas de error, medirlas, acotarlas y mostrarlas es la forma honesta de merecer la confianza del profesional.

8 Referencias

[1] Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., Ho, D. E. (2024). *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools*. Stanford RegLab / Yale; publicado en Journal of Empirical Legal Studies (2025). hai.stanford.edu/news/ai-trial-legal-models-hallucinate-1-out-6-or-more-benchmarking-queries

[2] Dahl, M., Magesh, V., Suzgun, M., Ho, D. E. (2024). *Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models*. Journal of Legal Analysis, 16(1). arxiv.org/abs/2401.01301

[3] *Mata v. Avianca, Inc.*, 678 F. Supp. 3d 443 (S.D.N.Y. 2023) — sanción por precedentes ficticios generados por IA.

[4] Wilson, E. B. (1927). *Probable Inference, the Law of Succession, and Statistical Inference*. Journal of the American Statistical Association, 22(158) — base del intervalo de confianza utilizado.

[5] Biblioteca del Congreso Nacional de Chile — fuente oficial de los textos normativos del corpus. bcn.cl/leychile

[6] AbogadoIA (2026). Reportes internos de evaluación: benchmark 27-feb-2026 (n=115) y re-validación 4-jul-2026 (arnés de evaluación permanente). Disponibles bajo acuerdo de confidencialidad para due diligence técnica.

> **Nota de confidencialidad técnica.** Este documento describe la arquitectura a nivel de diseño y su evaluación. Los componentes propietarios —la implementación del verificador de citas, los umbrales internos del puntaje de confianza, las estrategias de expansión de consulta, la identidad de los modelos empleados y los datasets completos de evaluación— no se divulgan aquí y están disponibles bajo acuerdo de confidencialidad para procesos de due diligence técnica.