INNOVA Y CREE · UNIDAD DE SISTEMAS DE IA VERIFICABLE

IA verificable para dominios regulados: arquitectura antialucinación con soberanía de datos y su evaluación empírica en producción

Documento técnico-científico. Valida el doble problema del despliegue de modelos de lenguaje en el sector legal y de recursos humanos —**alucinación** y **pérdida de control sobre datos confidenciales**—, presenta una arquitectura de recuperación aumentada con **verificación determinista de citas**, tres modelos de despliegue soberano de datos, y los resultados de un benchmark de 278 consultas bajo la rúbrica Stanford RegLab.

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

Elaborado por
Equipo de Ingeniería · Innova y Cree

Contacto
innovaycree.com · david@innovaycree.com

Fecha
Julio 2026

RESUMEN

Los asistentes de IA generativa de propósito general presentan dos fallas estructurales que impiden su adopción responsable en dominios regulados: (i) **alucinación** — Stanford RegLab midió tasas de 17% y 33% en herramientas legales comerciales que se vendían como "libres de alucinaciones" [1], y a julio de 2026 se documentan ~1.490 resoluciones judiciales en el mundo con citas fabricadas por IA [8], incluidas dos sanciones en tribunales chilenos en 2026 [9][10]; y (ii) **pérdida de soberanía sobre datos confidenciales**, agravada en Chile por la entrada en vigencia de la Ley 21.719 (dic. 2026, multas de hasta 20.000 UTM) [11] y en México por la LFPDPPP 2025 y los lineamientos de la Barra Mexicana de Abogados que obligan a informar al cliente cuando su información se comparte con sistemas de IA externos [12][13]. Este documento presenta una arquitectura de generación aumentada por recuperación (RAG) con **cuatro barreras antialucinación** —corpus estructurado y citable, recuperación híbrida multi-consulta, abstención explícita y verificación determinista de citas contra base de datos— y **tres modelos de despliegue** que mantienen los datos bajo control jurídico y físico del cliente. En evaluación sobre un despliegue en producción (corpus normativo chileno de 16 cuerpos legales y 1.067 artículos; n=278 consultas): **84,2% de respuestas exactas** (IC 95%: 77,7–89,0), **2,5% de alucinación** (IC 95%: 1,0–6,3) bajo la definición más exigente disponible, **0 fabricaciones en 120 pruebas hostiles** y **0 citas fantasma en 179 citas verificadas**. La mejora entre versiones del motor se validó con prueba pareada de McNemar (p=0,0004). Se declaran las limitaciones del método.

1 El problema, validado con evidencia

1.1 La alucinación no es anecdótica: está medida y sancionada

El estudio de referencia del campo (Magesh, Surani, Dahl, Suzgun, Manning y Ho, Stanford RegLab / Yale, *Journal of Empirical Legal Studies*, 2025) evaluó con 202 consultas preregistradas y calificación por expertos humanos las dos principales plataformas de investigación legal con IA del mundo: **Lexis+ AI produjo respuestas exactas en 65% de los casos con 17% de alucinación; Westlaw AI-Assisted Research, 42% y 33%** [1]. Ambas se comercializaban con afirmaciones de estar libres de alucinación. La conclusión no es que esas plataformas sean malas: es que **la alucinación es una propiedad estadística de los modelos de lenguaje, no un defecto que un proveedor pueda declarar resuelto**.

Las consecuencias ya son disciplinarias y económicas. Una base de datos académica documenta ~1.490 resoluciones judiciales con citas fabricadas por IA en más de 25 jurisdicciones [8]. Hitos: *Mata v. Avianca* (SDNY, 2023, seis precedentes ficticios, multa de USD 5.000); Morgan & Morgan (2025, ocho de nueve citas falsas generadas por su IA interna). En Chile, en 2026: el 2º Juzgado Civil de Concepción multó a un abogado por jurisprudencia inexistente generada con IA (feb. 2026) [9] y el Tribunal de Defensa de la Libre Competencia declaró inadmisible una demanda por citar sentencias falsas, con multa de 1 UTA (Rol C-547-26, mar. 2026) [10]. En México, el Semanario Judicial de la Federación fijó en agosto de 2025 el criterio de que la IA solo puede usarse en tareas auxiliares, con declaración expresa de la herramienta empleada [14].

1.2 El segundo dolor: "¿quién ve mis datos?"

Aun si la alucinación estuviera resuelta, el segundo bloqueo de adopción es la confidencialidad. Un despacho de abogados que sube expedientes reales a un chatbot de consumo enfrenta tres riesgos simultáneos:

- **Riesgo ético-profesional.** La ABA (Formal Opinion 512, 2024) exige consentimiento informado del cliente antes de introducir información confidencial en herramientas que aprenden de los datos [2]. La Barra Mexicana de Abogados publicó en octubre de 2025 los primeros lineamientos de IA de América Latina, con el deber de informar al cliente cuando su información se comparte con sistemas externos [13]. En Chile, el Código de Ética (arts. 7 y 46) prohíbe facilitar acceso a soportes electrónicos con información del cliente; la lectura doctrinal dominante encuadra allí el uso de IA de consumo. En México, la revelación de secreto profesional tiene además sanción penal (CPF arts. 210-211: prisión y suspensión profesional).
- **Riesgo regulatorio.** La Ley 21.719 chilena (vigencia 1 dic. 2026) impone al proveedor que trata datos por cuenta de otro un contrato de encargo (art. 15 bis), deber de secreto, medidas de seguridad y notificación de brechas, con multas de hasta 20.000 UTM y 2–4% de los ingresos anuales por reincidencia [11]. La LFPDPPP mexicana de 2025 contempla multas de hasta ~$37,5

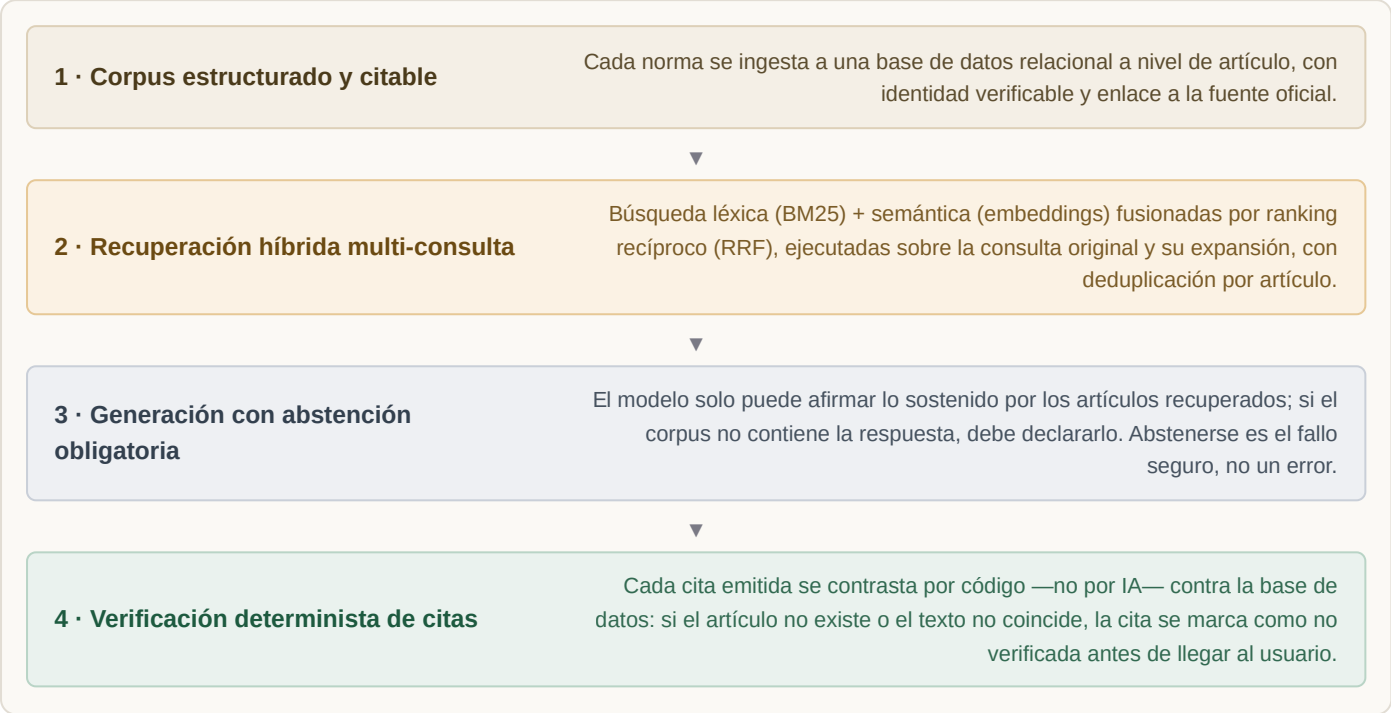
millones MXN, **duplicables cuando se trata de datos sensibles** [12] — y los datos de RRHH lo son con frecuencia: cuestionarios NOM-035 (salud), denuncias bajo reserva legal (Ley Karin en Chile), afiliación sindical y biometría.

- **Riesgo de gobernanza del dato.** Los planes de consumo de los chatbots no ofrecen garantías contractuales de no-retención ni de no-entrenamiento; los términos de servicio de varias herramientas permiten usar el contenido para mejorar los modelos. El episodio Samsung (2023, código fuente filtrado a un chatbot, prohibición corporativa total) sigue siendo el caso canónico [15].

> **Síntesis del dolor.** Las organizaciones de los sectores legal y RRHH no rechazan la IA: rechazan (a) sistemas cuya tasa de error nadie mide ni publica, y (b) arquitecturas donde sus datos salen de su control jurídico. Cualquier solución seria debe atacar **ambos** problemas a la vez, con evidencia verificable.

2 La solución: arquitectura antialucinación con verificación determinista

El principio de diseño es que **la confiabilidad no se pide al modelo: se impone desde la arquitectura**. El modelo de lenguaje es un componente reemplazable dentro de un pipeline donde cada afirmación debe sobrevivir a controles que no dependen de ningún modelo.



2.1 Por qué cada capa existe

Capa 1 — el corpus es la fuente de verdad, no la memoria del modelo. Los modelos de lenguaje "recuerdan" el derecho de forma difusa; por eso inventan artículos plausibles. Al estructurar el corpus a nivel de artículo en una base de datos relacional, toda respuesta puede exigir un ancla verificable: número de norma, número de artículo, texto oficial y enlace a la fuente pública (en Chile, la Biblioteca del Congreso Nacional). Lo que no está en el corpus, no existe para el sistema.

Capa 2 — la mayoría de las "alucinaciones" nacen como fallas de recuperación. Nuestra evidencia interna es contundente: al corregir defectos del motor de recuperación (no del modelo), la cobertura de contexto correcto pasó de 82,5% a 97,2% y la abstención innecesaria cayó de 16,3% a 5,1% **sin aumentar la alucinación** (§4). La combinación de canal léxico y semántico con fusión de rankings es hoy práctica de referencia porque cada canal falla de forma distinta: el léxico pierde sinónimos, el semántico pierde números de artículo y términos técnicos exactos.

Capa 3 — la abstención como contrato de comportamiento. El sistema distingue "no lo sé" de "no está en el corpus". Un asistente jurídico que responde siempre es más peligroso que uno que se abstiene: la métrica que optimizamos no es la tasa de respuesta, sino la tasa de respuesta *exacta y fundada*.

Capa 4 — el verificador no es un modelo. Esta es la barrera diferencial: un componente determinista (código convencional, auditable línea a línea) recalcula cada cita contra la base de datos. Un modelo de lenguaje no puede "convencer" al verificador: o el artículo existe con ese contenido, o la cita se degrada visiblemente. En 179 citas marcadas como verificadas durante la evaluación, las 179 apuntaban a artículos reales (§4.3).

2.2 Justificación del stack tecnológico

DECISIÓN	ALTERNATIVAS EVALUADAS	JUSTIFICACIÓN

Python en el motor de recuperación y evaluación	Node.js, Java	Ecosistema dominante en NLP y recuperación de información; las librerías de embeddings, evaluación estadística y procesamiento de texto legal maduran primero en Python; el código de benchmark es directamente auditable por terceros académicos.
PostgreSQL + extensión vectorial (pgvector)	Bases vectoriales dedicadas (Pinecone, Weaviate)	Un solo motor transaccional guarda artículos, embeddings e índices léxicos: el verificador determinista consulta la *misma* base que la recuperación, eliminando desincronizaciones. Software libre, auto-alojable en la infraestructura del cliente — requisito de los modelos de despliegue soberano (§3).
Búsqueda híbrida BM25 + embeddings con RRF	Solo semántica; solo léxica	Medido en nuestro propio arnés: la fusión por ranking recíproco superó a cada canal aislado; el canal léxico opera sobre la consulta original del usuario (la expansión automática degrada BM25).
Modelos de lenguaje frontier vía API con acuerdo de cero retención (ZDR)	Modelos abiertos locales	Para razonamiento jurídico, los modelos frontier mantienen una brecha de calidad relevante sobre los abiertos ejecutables en hardware de oficina [7]. El sistema es agnóstico del proveedor: la capa 4 no confía en ninguno. Donde el requisito de soberanía lo exige, se despliega variante local (§3, nivel 2).
Seudonimización previa de datos personales	Envío directo	Un filtro automático de reconocimiento de entidades reemplaza nombres, identificadores nacionales y datos de contacto por seudónimos reversibles *antes* de cualquier llamada a un modelo externo, tanto en la consulta como en los fragmentos recuperados.

3 Soberanía de datos: tres modelos de despliegue

La respuesta a "¿quién tiene acceso a mis datos?" no puede ser una promesa: debe ser una propiedad estructural del despliegue, respaldada por contrato. Los tres niveles comparten el mismo motor (§2) y el mismo marco contractual; difieren en dónde viven los datos.

NIVEL	ARQUITECTURA	QUÉ GARANTIZA	PERFIL DE CLIENTE
N1 · IA Privada Gestionada	Documentos, base de conocimiento y vectores en un servidor contratado **a nombre del cliente**. Innova y Cree administra la capa de infraestructura por acceso revocable. Consultas seudonimizadas hacia modelo frontier bajo acuerdo de cero retención de datos (ZDR) y cláusula de no-entrenamiento.	El cliente puede revocar el acceso del proveedor y exportar la totalidad de sus datos en cualquier momento. Ningún dato personal identificable sale del servidor. Cumple contrato de encargo (Ley 21.719 art. 15 bis / LFPDPPP).	Despachos y áreas de RRHH de 1 a 50 personas. Mejor equilibrio calidad-costo.
N2 · Servidor Soberano	Equipo físico instalado **en las oficinas del cliente**, con modelo de lenguaje local para el material más sensible y ruta híbrida seudonimizada para el resto. Mantenimiento remoto restringido a la capa de sistema, sin alcance sobre documentos.	Los datos designados como críticos no salen del edificio; el sistema opera incluso sin conexión. Precedente comercial validado en Europa para despachos de abogados [6].	Litigios de alta sensibilidad, denuncias bajo reserva legal (Ley Karin), investigaciones internas, banca.
N3 · Nube corporativa del cliente	Despliegue dentro de la cuenta cloud de la organización (p. ej. AWS Bedrock / Azure), donde el proveedor del modelo no retiene ni ve los datos, con residencia regional y aislamiento single-tenant.	Gobernanza TI corporativa: los datos nunca abandonan el perímetro cloud que la empresa ya audita; opción de computación confidencial (enclaves cifrados) donde ni el operador de la nube puede leer la inferencia.	Empresas reguladas con área de TI propia.

3.1 El marco contractual común

- **Contrato de encargo de tratamiento** conforme al art. 15 bis de la Ley 21.719 (Chile) y a la figura del encargado de la LFPDPPP (México): objeto, finalidad, tipos de datos, deber de secreto que subsiste al término, y prohibición expresa de uso para fines propios — **incluido el entrenamiento de modelos**.
- **Cadena de no-retención documentada**: los acuerdos de cero retención con los proveedores de modelos se incorporan al contrato, de modo que el cliente conoce y puede auditar la cadena completa, no solo al integrador.
- **Registro de accesos auditable** y notificación de incidentes en plazos comprometidos.
- **Reversibilidad total**: exportación de corpus, vectores y bitácoras en formatos abiertos, sin dependencia del proveedor (los componentes de almacenamiento son software libre).

> **Posición de mercado verificada.** A julio de 2026, ninguna de las plataformas legales con IA que operan en Chile y México ofrece despliegue on-premise ni en servidor del cliente; las garantías vigentes en el mercado se limitan a certificaciones y cláusulas de no-entrenamiento en nube del proveedor [5]. Los tres niveles aquí descritos no tienen equivalente empaquetado en la región.

4 Evaluación empírica: benchmark de 278 consultas

Medición del 10 de julio de 2026 sobre un despliegue del sistema **en producción**, con un corpus de normativa regulatoria chilena de **16 cuerpos legales y 1.067 artículos**. Metodología alineada con Stanford RegLab [1] y FACTS Grounding (Google DeepMind, 2025) [3].

4.1 Diseño

- **278 consultas** ejecutadas contra la plataforma en producción, dataset generado con semilla fija (reproducible entre corridas).
- **Separación de roles**: las preguntas factuales fueron generadas por un modelo *distinto* del evaluado, a partir del texto de artículos reales y **sin nombrar el artículo** — se mide recuperación, no eco. El juez que califica es también un modelo distinto del que responde (evita el sesgo de autopreferencia).
- **Cinco categorías**: factual · trampa (artículos inexistentes, verificados contra la base de datos) · fuera de corpus · adversarial (premisa falsa embebida) · fundamentación.
- **Evaluación determinista** (contraste directo contra la base de datos, sin juicio de ningún modelo) para las categorías trampa, fuera-de-corpus y adversarial; juez independiente bajo rúbrica Stanford para las factuales, con separación determinista de las abstenciones.

4.2 Rúbrica (Stanford RegLab)

VEREDICTO	DEFINICIÓN
Exacta	Correcta y sostenida por el texto oficial citado.
Incompleta	Correcta hasta donde llega, pero omite una condición esencial.
Se abstuvo	El sistema declara que el corpus no contiene la respuesta. **No es alucinación**: es el fallo seguro.
Incorrecta	Describe mal el derecho.
Mal fundada	La afirmación puede ser cierta, pero la cita entregada no la respalda.

> **Alucinación = incorrecta + mal fundada.** Es la definición de Stanford y es más exigente que "no inventó una norma": una cita real que no sostiene lo afirmado también cuenta como alucinación.

4.3 Resultados

84,2%
RESPUESTAS EXACTAS (IC 95%: 77,7–89,0)

2,5%
ALUCINACIÓN, RÚBRICA STANFORD (IC 95%: 1,0–6,3)

120/120
PRUEBAS HOSTILES SIN FABRICACIÓN (IC 95%: ≥96,9%)

0/179
CITAS FANTASMA ENTRE LAS CITAS VERIFICADAS

VEREDICTO	N	TASA	IC 95% (WILSON)	VERSIÓN PREVIA
Exacta	133	**84,2%**	77,7–89,0%	71,9%
Incompleta	13	8,2%	4,9–13,6%	9,2%
Abstención	8	5,1%	2,6–9,7%	16,3%
Incorrecta	3	1,9%	0,6–5,4%	1,3%
Mal fundada	1	0,6%	0,1–3,5%	1,3%

n = 158 respuestas factuales juzgadas bajo rúbrica; el resto del dataset corresponde a las categorías de evaluación determinista (tabla 4.4). La mejora respecto de la versión previa del motor proviene de correcciones *estructurales* de recuperación —no de ajustes por pregunta— y es estadísticamente significativa: **prueba pareada de McNemar sobre las mismas preguntas, 18 mejoraron / 2 empeoraron, p = 0,0004**. La reducción principal es la abstención sobre contenido que sí estaba en el corpus (16,3% → 5,1%), lograda **sin aumentar la tasa de alucinación** (2,6% → 2,5%).

4.4 Seguridad ante entradas hostiles (evaluación determinista)

PRUEBA	RESULTADO
Consultas por artículos inexistentes (verificados contra la BD)	60/60
Consultas por normas fuera del corpus	40/40
Consultas con premisa falsa embebida	20/20
Total: cero fabricaciones	**120/120**

Por la regla de tres, cero fallos en 120 intentos acota la tasa real de fabricación a ≤2,5% con 95% de confianza — el dato honesto no es "cero", es la cota. Adicionalmente, las 179 citas que el sistema marcó como verificadas fueron recalculadas contra la base de datos: 0 apuntan a artículos inexistentes.

4.5 Qué significa —y qué no— la tasa de alucinación

Los cuatro casos de alucinación de esta corrida comparten un patrón: el sistema respondió desde un **artículo vecino o una norma hermana** de contenido similar y describió mal un detalle (un plazo, el órgano competente). **En ningún caso inventó una norma o un artículo inexistente.** Lectura práctica: aproximadamente 1 de cada 40 consultas puede contener un error de matiz jurídico, detectable en segundos porque toda respuesta entrega sus citas verificadas con enlace a la fuente oficial. Por eso el sistema se diseña como **asistente del profesional, no como sustituto de su juicio**.

SISTEMA	RESPUESTAS EXACTAS	ALUCINACIÓN
Westlaw AI-Assisted Research [1]	42%	33%
Lexis+ AI [1]	65%	17%
Este sistema (v2, jul. 2026)	**84,2%**	**2,5%**

La comparación no es una equivalencia metodológica: aquellos sistemas operan sobre jurisprudencia estadounidense y este sobre normativa regulatoria chilena. Se ofrece como referencia de orden de magnitud.

> **Por qué no prometemos "cero alucinaciones".** Es estadísticamente indefendible (incluso 0 fallos en 250 pruebas deja un intervalo superior de ~1,5%) y comercialmente deshonesto: los sistemas que Stanford midió en 17% y 33% se vendían como "hallucination-free", y la FTC estadounidense ya sancionó afirmaciones de exactitud sin sustento [4]. Nuestro compromiso es el contrario: **publicar la tasa, su intervalo de confianza y sus limitaciones, y re-medirla trimestralmente con el mismo arnés**.

5 Limitaciones declaradas

1. El juez de la rúbrica es un modelo de lenguaje, no un abogado. Las abstenciones se separan con un filtro determinista, pero **falta la validación humana de una muestra** con reporte del acuerdo inter-anotador (kappa). Está comprometida como siguiente hito del programa de evaluación.
2. Una corrida por versión del motor. Los modelos no son deterministas: por eso las mejoras se validan con prueba pareada (McNemar) sobre las mismas preguntas, nunca comparando promedios entre corridas.
3. El corpus evaluado (16 normas, 1.067 artículos) es un dominio regulatorio acotado; los resultados no se extrapolan automáticamente a otros corpus. Cada nuevo dominio se mide con el mismo arnés antes de entrar en producción.
4. Persisten 8,2% de respuestas incompletas y 5,1% de abstenciones; su reducción es trabajo en curso y cada cambio se re-mide de forma pareada.
5. No existe aún un benchmark público independiente de IA jurídica en español; la comparación internacional (§4.5) es de orden de magnitud. La publicación periódica de este arnés busca contribuir a cerrar ese vacío.

6 Conclusión

El bloqueo de adopción de IA en los sectores legal y de RRHH es doble y está documentado: tasas de alucinación medidas de dos dígitos en los líderes del mercado, sanciones judiciales crecientes —ya también en Chile y con criterio judicial en México—, y un marco regulatorio (Ley 21.719, LFPDPPP 2025) que convierte el manejo informal de datos confidenciales en un riesgo económico y penal concreto. La respuesta de Innova y Cree ataca ambos frentes con propiedades estructurales, no promesas: una arquitectura donde toda cita se verifica por código contra la fuente, la abstención es el comportamiento por defecto ante la incertidumbre, y los datos permanecen bajo el control jurídico —o físico— del cliente en cualquiera de los tres modelos de despliegue. Los resultados (84,2% exactas; 2,5% de alucinación con su intervalo; 0 fabricaciones en 120 ataques; 0 citas fantasma) se publican con su metodología completa y sus limitaciones, y se re-miden trimestralmente. En un mercado donde nadie publica sus tasas de error, medirlas y mostrarlas es el diferencial.

7 Referencias

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

[2] American Bar Association (2024). *Formal Opinion 512 — Generative Artificial Intelligence Tools*. americanbar.org

[3] Google DeepMind (2025). *FACTS Grounding: benchmark de fundamentación factual*. deepmind.google

[4] Federal Trade Commission (2025). *FTC Order Requires Workado to Back AI Detection Claims*. ftc.gov/news-events/news/press-releases/2025/04

[5] Relevamiento propio de mercado legaltech Chile–México (jul. 2026): plataformas líderes regionales y su oferta de garantías de datos; ninguna con despliegue on-premise. Fuentes públicas de cada proveedor.

[6] Precedente comercial europeo de servidores de IA on-premise para despachos de abogados (Alemania, 2025–2026), comercializados bajo §43e BRAO / §203 StGB (secreto profesional).

[7] Comparativas públicas de modelos frontier vs. modelos abiertos ejecutables localmente (2026): brecha promedio superior a 25 puntos en razonamiento. artificialanalysis.ai

[8] Charlotin, D. (2026). *AI Hallucination Cases Database* (~1.490 resoluciones documentadas). damiencharlotin.com/hallucinations

[9] 2º Juzgado Civil de Concepción (feb. 2026): multa a abogado por jurisprudencia inexistente generada con IA. cnnchile.com

[10] Tribunal de Defensa de la Libre Competencia, Rol C-547-26 (mar. 2026): demanda inadmisible por sentencias falsas; multa 1 UTA. centrocompetencia.com/jurisprudencia-fantasma

[11] Ley 21.719 de Protección de Datos Personales, Chile (D.O. 13 dic. 2024, vigencia 1 dic. 2026). bcn.cl/leychile

[12] Ley Federal de Protección de Datos Personales en Posesión de los Particulares, México (DOF 20 mar. 2025).

[13] Barra Mexicana, Colegio de Abogados (oct. 2025). *Lineamientos para el Uso Responsable de la IA en el Ejercicio Profesional del Derecho*. bma.org.mx

[14] Semanario Judicial de la Federación (22 ago. 2025): criterio sobre uso auxiliar de IA por órganos jurisdiccionales.

[15] Forbes (may. 2023). *Samsung Bans ChatGPT After Sensitive Code Leak*.

> **Nota de confidencialidad técnica.** Este documento describe la arquitectura a nivel de diseño y su evaluación. Los componentes propietarios del sistema —la implementación del verificador, las estrategias de expansión de consulta, la ingeniería de contexto y los datasets completos de evaluación— no se divulgan en este documento y están disponibles bajo acuerdo de confidencialidad para procesos de due diligence técnica.