

Cómo se verifica una cita legal generada por IA

Verificar una cita legal generada por inteligencia artificial es comprobar cuatro cosas contra el texto oficial: que la norma exista, que el artículo citado sea el que se transcribe, que la versión esté vigente a una fecha y que lo citado respalde efectivamente la afirmación. Las cuatro se comprueban por código contra el corpus, no pidiéndole al modelo que se revise a sí mismo. Explicamos el método completo, la abstención como salida legítima, la tasa que arroja medido con su intervalo, los cinco límites que no resuelve y qué exigirle a un proveedor.

Por [David Pinto](#) — Fundador y arquitecto de la plataforma · [LinkedIn](#)

Revisión técnica: Innova y Cree

Publicado 17 ago 2026 · 14 min de lectura · [PDF](#)

Verificar una cita legal generada por inteligencia artificial consiste en comprobar cuatro cosas contra el texto oficial: que la norma **exista**, que el artículo citado sea el que se transcribe (**correspondencia**), que la versión invocada esté **vigente** a una fecha, y que lo citado **respalde** efectivamente la afirmación que acompaña. Las cuatro se comprueban por código contra el corpus indexado, no pidiéndole al modelo que se revise a sí mismo. Este artículo explica el método completo, por qué la abstención es una salida legítima del sistema, qué tasa arroja cuando se mide con método publicado, los cinco límites que no resuelve y qué exigirle a un proveedor para comprobar que existe. Es la contraparte operativa de [cómo medimos las alucinaciones de una IA en producción](#) y forma parte de la serie sobre [cómo evaluar a un proveedor de IA](#).

AVISO

Este artículo describe una práctica técnica de verificación documental. No constituye asesoría jurídica: la interpretación de una norma, su aplicación a un caso y la responsabilidad por lo que se presente ante un tribunal siguen siendo del profesional que firma. Ninguna cifra aquí publicada es una propiedad permanente de un sistema; toda tasa corresponde a una versión y una fecha concretas.

Qué significa «cita verificada» (definición operativa)

Una cita está verificada cuando un proceso independiente del modelo ha comprobado, contra el texto oficial, que la referencia existe, que su contenido es el que se transcribe, que la versión es la aplicable a la fecha consultada y que lo transcrito sostiene la afirmación a la que acompaña. Mientras falte cualquiera de esas cuatro comprobaciones, lo que hay es una cita **plausible**, que es otra cosa.

La distinción no es retórica. El estudio de Stanford RegLab y Yale sobre las principales herramientas de investigación jurídica con IA del mercado estadounidense fijó una definición operativa que conviene adoptar tal cual: una respuesta puede ser *incorrecta* —describe mal el derecho— o *mal fundada* —la afirmación podría ser cierta, pero la cita entregada no la respalda—, y ambas cuentan como alucinación¹. La segunda categoría es la que rompe la intuición del comprador: hay respuestas con citas reales, verificables una por una en el sitio oficial, que siguen siendo falsas porque el artículo citado no dice lo que la respuesta afirma que dice.

Las cuatro comprobaciones

Cada cita que produce el sistema pasa por cuatro pruebas, en este orden, y cualquiera que falle degrada la respuesta antes de que llegue al usuario.

Existencia. ¿La norma y el artículo referidos están en el corpus? Es una búsqueda exacta por identificador, no semántica: número de norma, número de artículo, inciso o fracción. Si el identificador no resuelve, la cita se descarta. Esta prueba es la que atrapa el caso clásico —el artículo inventado con numeración verosímil— y es trivial de implementar; su ausencia en un producto es señal de que nadie pensó el problema.

Correspondencia. ¿El texto que la respuesta presenta como cita coincide con el texto almacenado para ese identificador? Se compara el fragmento entregado contra el fragmento del corpus. No hace falta identidad literal absoluta —hay elisiones y puntos suspensivos legítimos—, pero sí que cada segmento citado aparezca en el original. Esto atrapa la mezcla de dos artículos en uno, que es un fallo frecuente y difícil de detectar leyendo.

Vigencia. ¿La versión citada es la aplicable a la fecha de la consulta? Un texto legal no es un documento sino una serie de versiones: la misma norma dice cosas distintas según el día. El corpus debe almacenar la versión con su fecha, y la respuesta debe declarar a qué fecha corresponde lo que afirma. Los repositorios oficiales trabajan así: LeyChile publica cada norma con su fecha de versión y su estado de derogación⁸, el Boletín Oficial del Estado español publica textos consolidados con fecha de actualización¹⁰ y la compilación de la Cámara de Diputados de México indica la última reforma incorporada⁹. Un sistema que no guarda la fecha de versión no puede responder esta pregunta, y por tanto no puede verificar nada.

Suficiencia. ¿Lo citado respalda la afirmación? Es la comprobación más costosa y la que más fallos revela. Consiste en contrastar cada aserción de la respuesta contra el fragmento que la sostiene, y marcar la que no se sostiene. FACTS Grounding, el banco de pruebas publicado por Google DeepMind, formaliza exactamente esta idea: descartar primero las respuestas que eluden la pregunta —para que abstenerse de todo no sea una forma de ganar— y verificar después, frase por frase, que nada afirme lo que el documento entregado no dice³.

Por qué el modelo no debe citar de memoria

Un modelo de lenguaje produce texto estadísticamente plausible; su seguridad al expresarse es independiente de su exactitud. El NIST nombra este comportamiento **confabulación** —contenido falso o engañoso producido con confianza— y lo incluye entre los riesgos propios de la IA generativa, con la instrucción de medirlo, no de negarlo⁴. OWASP lo recoge en su lista de riesgos para aplicaciones con modelos de lenguaje bajo la etiqueta de desinformación⁶.

Por eso la arquitectura correcta no le pregunta al modelo qué dice la ley: le entrega los fragmentos del corpus pertinentes a la consulta y le pide que responda **solo** con eso, citando el identificador de cada fragmento. La diferencia práctica es que el espacio de respuestas queda acotado por documentos reales. La diferencia medida es sustancial pero no total: el trabajo previo del mismo grupo de Stanford mostró que los modelos generalistas alucinan en más de la mitad de las consultas jurídicas directas², y el estudio de 2025 mostró que recuperar documentos reduce el problema sin eliminarlo¹. La recuperación es condición necesaria; la verificación posterior es lo que convierte una reducción en una garantía comprobable.

Hay una consecuencia de diseño que conviene explicitar: los fragmentos que viajan al modelo son solo los pertinentes a cada consulta, nunca el corpus completo ni el expediente entero. Eso importa para la exactitud y, además, para la exposición de datos; el detalle de qué componente ve qué se declara en el [inventario de componentes de IA](#).

La abstención es una salida del sistema, no un fallo

Un sistema serio tiene una respuesta prevista para «no hay fundamento en el corpus»: decirlo, decir qué se buscó y no ofrecer nada más. Esa salida se llama abstención y es una decisión de producto, no una avería.

Los modelos generalistas están afinados para ser útiles, y responder siempre es la conducta que optimiza esa utilidad aparente. En un dominio donde una afirmación falsa se convierte en un escrito presentado ante un tribunal, esa utilidad es el problema. La definición operativa que usamos separa tres cosas que suelen confundirse: **alucinar** (afirmar algo incorrecto o mal fundado), **abstenerse** (declarar que no hay fundamento) y **quedarse corto** (una respuesta correcta que omite una condición esencial). Solo la primera cuenta como alucinación; la segunda y la tercera se reportan aparte, porque un sistema que se abstiene demasiado es inútil aunque nunca se equivoque.

La comprobación es sencilla para un comprador: pedir una consulta cuya respuesta correcta sea «eso no existe». Si el sistema produce una respuesta bien redactada en lugar de abstenerse, el resto de las garantías no importa.

La verificación se ejecuta por código, no con un segundo modelo

Las comprobaciones de existencia, correspondencia y vigencia son deterministas: son consultas al corpus y comparaciones de texto. Ejecutarlas por código tiene tres ventajas sobre pedirle a un modelo que revise: es reproducible, es auditable línea por línea y cuesta esencialmente nada por consulta.

El motivo de fondo es otro. Un modelo que evalúa respuestas —propias o ajenas— arrastra sesgos documentados de posición, de verbosidad y de autopreferencia; por eso FACTS Grounding promedia el juicio de varios modelos de distintos proveedores en lugar de confiar en uno³. Trasladado a producción: el juez no puede ser el evaluado. Cuando la comprobación exige criterio —la suficiencia—, el evaluador debe ser un componente distinto del que generó la respuesta, y su calibración debe contrastarse contra revisión humana sobre una muestra. Todo lo anterior es la función MEASURE del marco de gestión de riesgo del NIST aplicada a un caso concreto: métodos documentados, resultados registrados, no impresiones⁵.

Dónde se comprueba una cita: la fuente oficial por país

La verificación termina siempre en el mismo lugar: el repositorio oficial de la jurisdicción. Un sistema que se vende en varios países necesita una fuente primaria por cada uno, con su propia forma de expresar versiones y vigencias. Esta tabla reúne las que usamos como referencia de contraste, consultadas el 17 de agosto de 2026.

JURISDICCIÓN	REPOSITORIO OFICIAL DE REFERENCIA	CÓMO EXPRESA LA VIGENCIA
Chile	LeyChile, Biblioteca del Congreso Nacional	Versión con fecha y estado de derogación por norma ⁸
México	Compilación de textos vigentes de la Cámara de Diputados, sobre lo publicado en el Diario Oficial de la Federación	«Última reforma DOF» con fecha en la portada de cada ley ⁹
España y UE	Boletín Oficial del Estado (textos consolidados) y EUR-Lex	Texto consolidado con fecha de actualización y bloques modificados
Colombia	Diario Oficial y compilaciones oficiales del Congreso	Notas de vigencia y jurisprudencia concordante por artículo
Perú	El Peruano y el Sistema Peruano de Información Jurídica	Norma con fecha de publicación y modificatorias
Argentina	Boletín Oficial e InfoLEG	Texto actualizado con las normas modificatorias enlazadas

repositorios oficiales usados como fuente de contraste para verificar una cita, por jurisdicción.

AVISO

Comprobamos el contenido de los repositorios de Chile, México y España desde nuestra propia infraestructura el 17 de agosto de 2026, incluida la lectura del servicio de intercambio de normas de LeyChile para confirmar fecha de versión y estado de derogación. Los repositorios de Colombia, Perú y Argentina se citan por su nombre oficial sin enlace directo porque no pudimos confirmar por contenido una URL estable desde nuestro entorno; verificar la URL vigente antes de apoyarse en ella. Un código de respuesta correcto no prueba que la página sirva el documento esperado: esa confusión es un error que hemos cometido y corregido.

Qué arroja el método cuando se mide

El resultado de todo lo anterior no es una promesa: es una tasa, con fecha, versión del sistema e intervalo de confianza. Publicamos las nuestras congeladas por medición, sin actualizarlas con el corpus vivo⁷.

MEDICIÓN DEL 10 JUL 2026

278 consultas contra un sistema en producción sobre corpus normativo: 84,2 % de respuestas exactas y una tasa de alucinación de 2,5 % (incorrectas más mal fundadas), con intervalos de Wilson al 95 % publicados. Fuente: Innova y Cree, /empresas/benchmark/, consultado el 17 de agosto de 2026.

Sobre las pruebas diseñadas para inducir el error —artículos inexistentes, normas fuera del corpus, preguntas con premisa falsa— la lectura honesta no es un absoluto sino una cota. Con 120 pruebas hostiles y 0 fabricaciones observadas, lo que puede afirmarse es que la tasa real está por debajo de aproximadamente tres dividido por el número de pruebas, no que sea nula: ningún número de ensayos sin fallo demuestra la imposibilidad del fallo. La misma regla aplica a la batería interna sobre corpus legal chileno, con 115 consultas y una cota superior de 3,5 %.

COTA, NO ABSOLUTO

115 consultas en la batería interna sobre corpus legal chileno, sin que se observara ninguna respuesta alucinada; la lectura estadística correcta es una cota superior de 3,5 % al 95 % de confianza. Fuente: Innova y Cree, tabla maestra de cifras, corte del 13 de agosto de 2026.

Como orden de magnitud externo, el estudio de Stanford RegLab midió 33 % y 17 % de alucinación en dos plataformas comerciales que se anunciaban libres de ese defecto, sobre 202 consultas evaluadas por expertos¹. No son cifras comparables entre sí —distinto corpus, distinta jurisdicción, distinto tipo de pregunta—; lo comparable es el método.

Los cinco límites del método

Un método serio se describe por lo que no cubre tanto como por lo que cubre.

- 1. Corpus incompleto.** La verificación solo alcanza lo indexado. Si una norma aplicable no está en el corpus, el sistema no la citará mal: simplemente no la citará, y la respuesta será incompleta sin que ninguna comprobación lo señale. Por eso la cobertura declarada del corpus, con su fecha de corte, es parte de la ficha del producto.
- 2. Desfase de vigencia.** Entre la publicación de una reforma y su incorporación al corpus hay una ventana. El sistema debe declarar la fecha de corte de cada fuente y, en consultas sensibles a cambios recientes, advertirlo.
- 3. Interpretación.** Que una cita sea exacta no hace correcta la conclusión jurídica. Dos artículos correctamente citados pueden sostener lecturas opuestas; elegir entre ellas es trabajo profesional y no se automatiza.
- 4. Jurisdicción cruzada.** El error más silencioso es citar bien una norma del país equivocado. La comprobación de existencia debe estar acotada al corpus de la

jurisdicción de la consulta, no al corpus completo.

5. **Responsabilidad.** Quien firma el escrito responde por él. Ningún sistema traslada esa responsabilidad, y ningún proveedor serio ofrece hacerlo.

Cómo exigirlo a un proveedor: cinco pruebas de una hora

Estas cinco pruebas se ejecutan en una sesión con el sistema delante y separan un producto verificable de una demostración ensayada.

#	QUÉ PEDIR	QUÉ DEBE PASAR
1	Una respuesta real con el enlace al texto oficial de cada cita	Cada enlace abre la norma y el artículo exactos, en el repositorio oficial, no en una copia del proveedor
2	Una consulta cuya respuesta correcta sea «esa norma no existe»	El sistema se abstiene y explica qué buscó; no redacta una respuesta plausible
3	Una consulta sobre una norma reformada recientemente	La respuesta declara la fecha de versión y la fecha de corte del corpus
4	La tasa medida, con fecha, versión evaluada e intervalo	Hay número, método y fecha; si la respuesta es un adjetivo, no hay medición
5	La cobertura del corpus y qué queda fuera	Existe una lista de fuentes con fecha de corte y una declaración de lo no cubierto

cinco comprobaciones que un comprador puede ejecutar en una sola sesión técnica.

Un proveedor que responde con evidencia a las cinco está describiendo un sistema; uno que responde con adjetivos está describiendo una expectativa. La forma de dejarlo por escrito —umbrales, re-medición ante cambio de modelo, aviso previo— pertenece a la [política de IA responsable](#), y el protocolo completo de medición, con el

arnés publicado, está en [cómo medimos las alucinaciones de una IA en producción](#). Si quiere aplicar estas pruebas a un proveedor concreto, puede [plantearnos el caso](#).

1. Magesh, Surani, Dahl, Suzgun, Manning y Ho, *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools*, *Journal of Empirical Legal Studies* 22(2), 2025 (arXiv 2405.20362). Consultado el 17 de agosto de 2026. [↩↩↩](#)
2. Dahl, Magesh, Suzgun y Ho, *Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models*, *Journal of Legal Analysis* 16(1), 2024 (arXiv 2401.01301). Consultado el 17 de agosto de 2026. [↩](#)
3. Google DeepMind, *FACTS Grounding: A new benchmark for evaluating the factuality of large language models*, 17 de diciembre de 2024. Consultado el 17 de agosto de 2026. [↩↩](#)
4. NIST, *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile* (NIST AI 600-1), julio de 2024; la confabulación figura entre los riesgos propios de la IA generativa. Consultado el 17 de agosto de 2026. [↩](#)
5. NIST, *AI Risk Management Framework 1.0*, 26 de enero de 2023, función MEASURE. Consultado el 17 de agosto de 2026. [↩](#)
6. OWASP GenAI Security Project, *LLM Top 10 — 2026*. Consultado el 17 de agosto de 2026. [↩](#)
7. Innova y Cree, *Benchmark anti-alucinación*, con metodología, intervalos y fecha de medición. Consultado el 17 de agosto de 2026. [↩](#)
8. Biblioteca del Congreso Nacional de Chile, LeyChile. Verificado el 17 de agosto de 2026 contra su servicio XML de intercambio de normas, que devuelve fecha de versión, fecha de publicación y estado de derogación. [↩↩](#)
9. Cámara de Diputados del H. Congreso de la Unión (México), *Código Fiscal de la Federación*, texto vigente compilado; el artículo 29 regula los comprobantes fiscales digitales por Internet. Consultado el 17 de agosto de 2026. [↩↩](#)

10. Boletín Oficial del Estado (España), texto consolidado de la Ley 18/2022, citado aquí como ejemplo de consolidación con fecha. Consultado el 17 de agosto de 2026. [↩](#)

Puntos clave

- 01 Una cita está verificada cuando pasa cuatro comprobaciones contra el texto oficial: existencia, correspondencia, vigencia a una fecha y suficiencia respecto de lo afirmado.
- 02 Las cuatro se ejecutan por código contra el corpus; pedirle al modelo que revise sus propias citas reproduce el sesgo de autopreferencia y no sirve como control.
- 03 La abstención explícita es una salida del sistema, no una avería: se cuenta aparte de la alucinación y su ausencia es señal de que el sistema fue configurado para responder siempre.
- 04 El resultado se expresa como tasa medida con fecha, versión e intervalo de confianza; «no inventa» no es una propiedad, es el resumen impreciso de una medición.
- 05 El método no resuelve cinco cosas: corpus incompleto, desfase de vigencia, interpretación, jurisdicción cruzada y la responsabilidad profesional de quien firma.

Preguntas frecuentes

¿Una IA con citas puede seguir inventando?

¿Qué diferencia hay entre verificar la cita y verificar la respuesta?

¿Por qué no basta con pedirle al modelo que revise sus propias citas?

¿Qué debe hacer el sistema cuando no encuentra fundamento?

Referencias

- [1] Magesh, Surani, Dahl, Suzgun, Manning, Ho — «Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools», Journal of Empirical Legal Studies 22(2), 2025 (arXiv 2405.20362). <https://arxiv.org/abs/2405.20362>. Consultado el 17 de agosto de 2026. PAPER
- [2] Dahl, Magesh, Suzgun, Ho — «Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models», Journal of Legal Analysis 16(1), 2024 (arXiv 2401.01301). <https://arxiv.org/abs/2401.01301>. Consultado el 17 de agosto de 2026. PAPER
- [3] Google DeepMind — «FACTS Grounding: A new benchmark for evaluating the factuality of large language models» (17 dic 2024). <https://deepmind.google/discover/blog/facts-grounding-a-new-benchmark-for-evaluating-the-factuality-of-large-language-models/>. Consultado el 17 de agosto de 2026. PAPER
- [4] NIST AI 600-1 — Generative AI Profile: la «confabulación» como riesgo propio de la IA generativa. <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf>. Consultado el 17 de agosto de 2026. NORMA
- [5] NIST AI Risk Management Framework 1.0 (función MEASURE). <https://www.nist.gov/itl/ai-risk-management-framework>. Consultado el 17 de agosto de 2026. NORMA
- [6] OWASP GenAI — LLM Top 10 (edición 2026): desinformación y contenido no fundamentado. <https://genai.owasp.org/resource/owasp-genai-llm-top-10-2026/>. Consultado el 17 de agosto de 2026. NORMA
- [7] Innova y Cree — Benchmark anti-alucinación: tasas medidas con su metodología, intervalos y fecha de corte. <https://innovaycree.com/empresas/benchmark/>. Consultado el 17 de agosto de 2026. PROPIO
- [8] Biblioteca del Congreso Nacional de Chile — LeyChile (texto oficial y versiones por fecha; verificado contra su servicio XML de intercambio de normas). <https://www.bcn.cl/leychile/navegar?idNorma=1209272>. Consultado el 17 de agosto de 2026.

LEY

[9] Cámara de Diputados (México) — Código Fiscal de la Federación, texto vigente compilado (art. 29: comprobantes fiscales digitales por Internet).

<https://www.diputados.gob.mx/LeyesBiblio/pdf/CFF.pdf>. Consultado el 17 de agosto de 2026.

LEY

[10] Boletín Oficial del Estado (España) — texto consolidado de la Ley 18/2022, ejemplo de versión consolidada con fecha. <https://www.boe.es/buscar/act.php?id=BOE-A-2022-15818>.

Consultado el 17 de agosto de 2026. LEY

Citar este artículo

APA

Pinto, D. (2026, 17 de agosto). Cómo se verifica una cita legal generada por IA (v1.0). Blog técnico de Innova y Cree. <https://innovaycree.com/blog/ia-verificable/como-verificamos-cada-cita.html>

BIBTEX

```
@misc{pinto2026como,
  author      = {Pinto, David},
  title       = {Cómo se verifica una cita legal generada por IA},
  howpublished = {Blog técnico de Innova y Cree},
  year        = {2026},
  month       = {ago},
  note        = {v1.0, revisado 2026-08-17},
  url         = {https://innovaycree.com/blog/ia-verificable/como-verificamos-cada-cita.html}
}
```

v1.0 · 17 ago 2026

► HISTORIAL DE VERSIONES

SOBRE EL AUTOR

[David Pinto](#) · Fundador y arquitecto de la plataforma

Diseña la arquitectura de IA verificable de Innova y Cree: RAG con cita verificada contra la fuente oficial, benchmark anti-alucinación público y plataformas en producción para sectores regulados de habla hispana.

[LinkedIn](#) · [Todos sus artículos](#)