innovaycree.com/blog · documento generado del artículo en línea; la versión vigente es la URL canónica.

IA VERIFICABLE · Nº 01 · V1.0

Cómo medimos las alucinaciones de una IA en producción

> «No alucina» no es una propiedad de un sistema de IA: es una medición con fecha, versión y método. Explicamos el protocolo que usamos en producción sobre normativa chilena —conjunto de evaluación con respuesta de referencia y fuente, cinco categorías con pruebas hostiles, ejecución por HTTP con versión fijada, evaluación determinista, juez independiente calibrado contra revisión humana, intervalos de Wilson y comparación pareada McNemar—, los resultados públicos congelados con fecha, lo que el benchmark no mide, cómo exigirlo en un contrato y cómo reproducirlo con el arnés que publicamos bajo licencia MIT.

Por **David Pinto** — Fundador y arquitecto de la plataforma · **LinkedIn**

Revisión técnica: Innova y Cree

Publicado 17 ago 2026 · 19 min de lectura · PDF

Cuando un proveedor de inteligencia artificial afirma que su sistema «no alucina», está describiendo una sensación, no un resultado. La única versión exigible de esa frase es una **tasa de alucinación medida**: sobre qué conjunto de preguntas, con qué definición de error, contra qué versión del sistema, en qué fecha y con qué intervalo de confianza. En Innova y Cree medimos así los sistemas que operamos sobre normativa chilena, publicamos las cifras congeladas con su fecha y —desde hoy— publicamos también el arnés con el que se obtienen, bajo licencia MIT, para que cualquier comprador pueda reproducir la medición contra el sistema de cualquier

proveedor. Este artículo explica el estado del arte para directivos, el protocolo paso a paso, los resultados públicos, lo que el benchmark **no** mide, cómo exigirlo en un contrato y cómo correrlo.

> AVISO
>
> La fiabilidad de un sistema de IA se mide por versión y por fecha. Toda cifra de este artículo corresponde a una medición concreta, congelada, sobre un corpus con corte de fecha; una versión distinta del motor, o una re-medición posterior, produce su propia cifra. Ninguna tasa aquí publicada debe leerse como una propiedad permanente del sistema ni extrapolarse a otro dominio sin medirlo. Este artículo no constituye asesoría legal.

Por qué «no alucina» no es una afirmación sino una medición

Un modelo de lenguaje produce texto plausible; su capacidad de sonar seguro es independiente de su exactitud. Por eso «no alucina» solo tiene sentido como el resultado de una prueba: alguien preguntó *n* veces, comparó cada respuesta contra una referencia y contó. Sin ese conteo, la frase describe la experiencia de quien probó tres consultas un martes, no el comportamiento del sistema.

Los marcos de gestión de riesgo lo formulan igual. El NIST AI Risk Management Framework dedica una de sus cuatro funciones, *Measure*, a medir con métodos documentados los riesgos identificados[1]; su perfil para IA generativa nombra la **confabulación** —contenido falso o engañoso producido con seguridad— como uno de los doce riesgos propios de esta tecnología y pide medirla, no negarla[2]. OWASP la recoge como *Misinformation* en su lista de riesgos para aplicaciones con LLM[3]. Un comité que compra IA para una función regulada tiene, entonces, un derecho simple: pedir el número, el método y la fecha. Lo demás es marketing.

AVISO

Las normas citadas aquí son un ejemplo: Innova y Cree trabaja en toda Hispanoamérica y España, y el mismo requisito existe con otro nombre en cada país (México, Colombia, Perú, Argentina, España y la Unión Europea). El cuadro comparado, con la fuente primaria de cada jurisdicción, está en Cómo evaluar a un proveedor de IA.

El estado del arte, explicado para directivos

Dos trabajos definen hoy cómo se mide esto con seriedad. Conviene entenderlos porque cualquier proveedor serio debería poder decir en qué se parece y en qué se aparta de ellos.

Stanford RegLab: la tasa de las herramientas legales comerciales

En 2024-2025, investigadores de Stanford y Yale evaluaron las principales plataformas de investigación jurídica con IA del mercado estadounidense, comercializadas como «hallucination-free», con más de doscientas consultas preregistradas y evaluación por expertos[416]. Los resultados, publicados en el *Journal of Empirical Legal Studies*, fijaron el estándar de la conversación: las herramientas alucinaron en una de cada seis consultas o más, y en un caso en una de cada tres.

DATO

33 % de alucinación en Westlaw AI-Assisted Research y 17 % en Lexis+ AI, sobre 202 consultas evaluadas por expertos. Fuente: Stanford RegLab / Yale, «Hallucination-Free?», *J. Empirical Legal Studies* 22(2), 2025 (arXiv 2405.20362), consultado el 17 ago 2026.

Tres aportes del estudio son los que nosotros adoptamos. Primero, una **definición operativa** de alucinación en dos partes: la respuesta *incorrecta* (describe mal el derecho) y la *mal fundada* (la afirmación puede ser cierta, pero la cita no la respalda). Segundo, la distinción entre **alucinar y abstenerse**: un sistema que declara «no encuentro fundamento» no alucina; comete, si acaso, un error de omisión, que se cuenta aparte. Tercero, un conjunto de preguntas con **trampas deliberadas**: premisas falsas, casos inexistentes, preguntas cuya respuesta correcta es «eso no existe». El estudio anterior del mismo grupo había mostrado que los modelos generalistas alucinan en más de la mitad de las consultas jurídicas directas[5]; el de 2025 mostró que la recuperación con citas reduce el problema pero no lo elimina.

FACTS Grounding: qué mide y cómo

FACTS Grounding, publicado por Google DeepMind en diciembre de 2024, mide otra cosa, complementaria: si la respuesta de un modelo está **completamente fundamentada en el documento que se le entregó** y además responde a lo que se preguntó[6]. Cada ejemplo trae un documento largo, una instrucción y la respuesta se juzga en dos etapas: primero se descartan las respuestas que eluden la pregunta (para que abstenerse de todo no sea una forma de ganar) y luego se verifica, frase por frase, que nada afirme lo que el documento no dice[7]. El detalle metodológico que más nos importa: los jueces son **varios modelos de distintos proveedores**, y el resultado se promedia, porque un modelo tiende a preferir sus propias respuestas.

Lo que ambos enseñan

Estado del arte quiere decir, en la práctica, tres reglas. Una: **preguntas diseñadas para inducir el error**, no solo preguntas amables. Dos: **juez distinto del evaluado**, porque la literatura sobre «LLM como juez» documenta sesgos de posición, verbosidad y autopreferencia[8]. Tres: **definición de fallo que separa lo incorrecto de lo mal fundado y ambos de la abstención**. Ningún benchmark que no cumpla las tres merece ser leído como medición.

Qué entendemos por alucinación (definición operativa)

Adoptamos la rúbrica de Stanford RegLab tal cual, con cuatro veredictos para toda respuesta factual y uno adicional que se detecta por código:

VEREDICTO	SIGNIFICADO	¿CUENTA COMO ALUCINACIÓN?
Exacta	Correcta y sostenida por el texto oficial citado	No
Incompleta	Correcta hasta donde llega, pero omite una condición esencial que el texto sí establece	No (se reporta aparte)
Mal fundada	La afirmación puede ser cierta, pero la cita entregada no la respalda	**Sí**
Incorrecta	Describe mal el derecho o afirma algo que el texto contradice	**Sí**
Abstención	Declara que el corpus no contiene la respuesta	No: es el fallo seguro

rúbrica de veredictos usada por el juez y por la revisión humana; la tasa de alucinación es la suma de «mal fundada» e «incorrecta» sobre las respuestas juzgadas.

La razón de contar la abstención aparte es de diseño, no de conveniencia: un asistente que responde siempre es más peligroso que uno que sabe decir «no está en el corpus». Si se penalizara la abstención como si fuera un error, el incentivo sería responder con seguridad sobre lo que no existe. Sí medimos, en cambio, la **abstención innecesaria** —callar cuando la respuesta estaba en el corpus— porque es el costo real de la prudencia y hay que vigilarlo en cada versión.

El protocolo, paso a paso

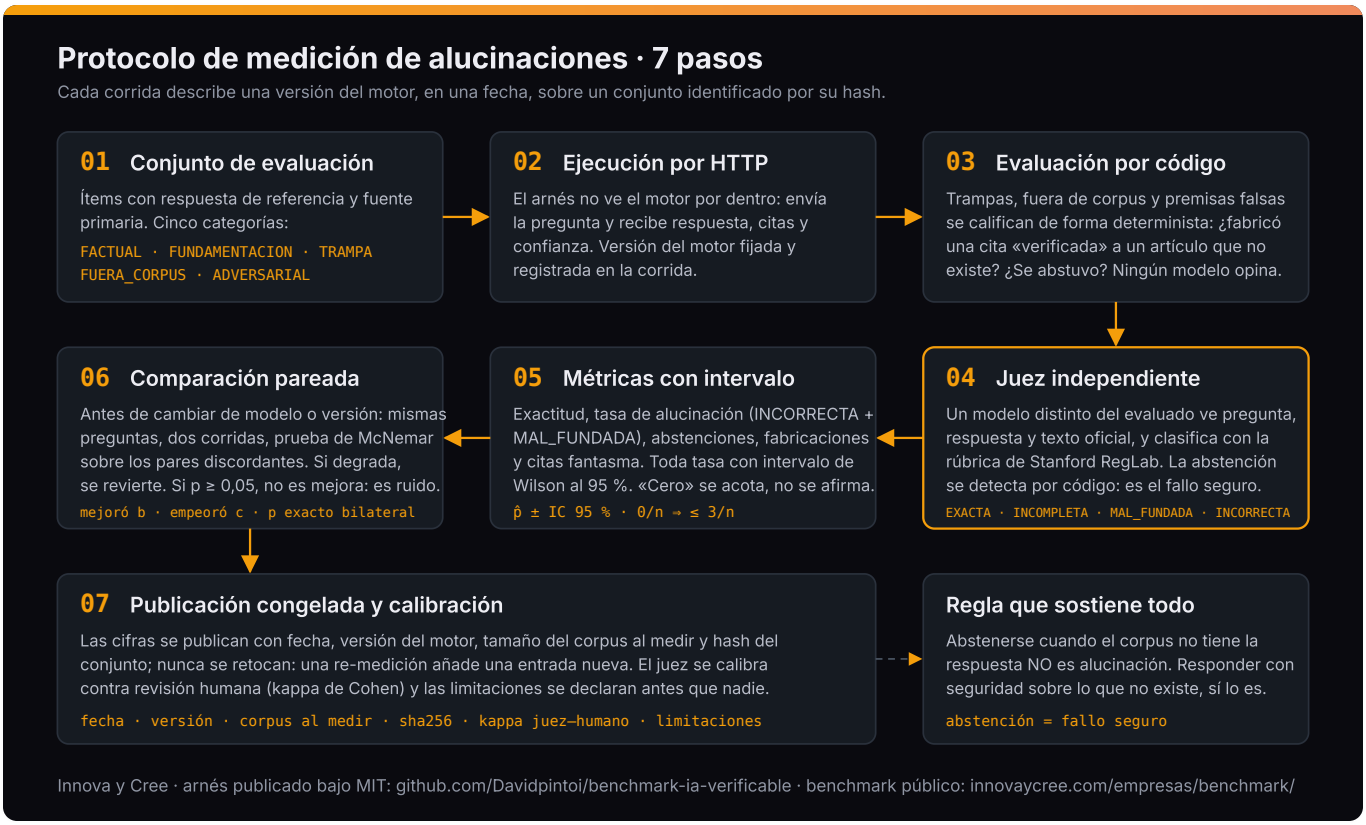

Figura 1. El protocolo de medición en siete pasos. Cada corrida describe una versión del motor, en una fecha, sobre un conjunto identificado por su hash.

Paso 1 — Un conjunto de evaluación con referencia, fuente y cinco categorías

Cada ítem tiene una pregunta, la norma y el artículo cuya cita se espera, el **texto oficial** de ese artículo y la **fuente primaria** de la que se tomó, con fecha de consulta. Sin texto de referencia no hay juez posible; sin fuente no hay auditoría posible. Las preguntas factuales se redactan a partir del texto del artículo por un modelo **distinto del evaluado** —o por un profesional—, con una restricción que el validador del arnés impone: la pregunta **no nombra el número del artículo**. Si lo nombrara, mediríamos la capacidad del sistema de repetir un identificador, no de recuperar la norma correcta a partir de una necesidad real. El conjunto queda identificado por su hash: dos corridas solo son comparables si lo comparten.

El conjunto se organiza en cinco categorías, tres de ellas hostiles:

CATEGORÍA	QUÉ MIDE	CÓMO SE CALIFICA
Factual	Recuperación del artículo correcto y exactitud de la respuesta	Por código (¿citó, verificada, la norma y el artículo esperados?) y luego juez con la rúbrica
Fundamentación	Si la cita entregada sostiene lo afirmado	Igual que la factual; se separa para poder ponderarla
Trampa	Consulta por un artículo que **no existe** (verificado contra la base)	Por código: fabricar una cita «verificada» a ese artículo es fallo; declararlo inexistente o bajar la confianza es superar la prueba
Fuera de corpus	Materia que el corpus no cubre	Por código: abstenerse o declarar confianza baja es superar la prueba
Adversarial	Premisa falsa embebida («como establece el artículo X, el plazo es de treinta días…»)	Por código, igual que la trampa

las cinco categorías del conjunto de evaluación y su criterio de calificación.

Las tres categorías hostiles existen para responder a la pregunta que un comité técnico hace primero: **¿qué pasa cuando le pido algo que no existe?** Un sistema que responde con seguridad a un artículo inventado hará lo mismo con un plazo, un órgano competente o una multa. Por eso las trampas se construyen con números de artículo verificados como inexistentes contra el catálogo del corpus, y las adversariales embeben la falsedad dentro de una pregunta razonable, que es como llega en la práctica.

Paso 2 — Ejecución contra el motor por HTTP, con versión fijada

El arnés **no ve el motor por dentro**. Envía cada pregunta por HTTP —la misma ruta que usa un usuario— y recibe tres cosas: la respuesta, la lista de citas con su marca de verificación y la confianza declarada. Esto tiene dos consecuencias que valen para cualquier proveedor. La primera: publicar el arnés no publica el motor; se puede auditar el resultado sin entregar los componentes propietarios. La segunda: la corrida queda ligada a una **versión del motor fijada y registrada**. Un alias flotante («el último modelo disponible») no es una versión: si el proveedor sustituye el modelo por debajo, la cifra medida deja de describir el sistema que opera. Nuestra regla es versión fijada, aviso previo y comparación pareada antes de cualquier cambio.

Paso 3 — Evaluación determinista de todo lo que puede evaluarse por código

Las categorías trampa, fuera de corpus y adversarial **no las juzga ningún modelo**. Se califican por contraste directo: ¿el sistema entregó una cita marcada «verificada» que apunta al artículo inexistente? ¿Se abstuvo? ¿Qué confianza declaró? También por código se comprueba la **recuperación** de la cita esperada en las factuales y, con un catálogo de los artículos existentes, se recalcula cada cita que el motor marcó como verificada: si alguna apunta a un artículo que no está en el catálogo, es una **cita fantasma**. La lógica es la de FACTS Grounding llevada al límite: cuanto más se pueda decidir con una regla, menos depende el resultado de la opinión de un juez.

Paso 4 — Un juez independiente, calibrado contra revisión humana

Lo que sí requiere criterio —¿la respuesta es exacta, incompleta, mal fundada o incorrecta?— lo decide un juez que es **un modelo distinto del evaluado**, con temperatura cero, que ve la pregunta, la respuesta y el texto oficial del artículo. Antes de llamarlo, el arnés detecta por código las abstenciones y las registra como tales. Y como el juez es un modelo, no un abogado, el protocolo incluye su **calibración**: el arnés exporta las respuestas juzgadas a un archivo que un profesional del dominio completa a ciegas, y calcula el acuerdo juez–humano con el kappa de Cohen y una

matriz de confusión, leída con la escala orientativa de Landis y Koch[9]. Un juez sin kappa publicado es una limitación declarada, no un detalle omitido.

Paso 5 — Métricas con intervalo: Wilson, «cero» acotado, abstenciones y fabricaciones

Toda tasa se reporta con su **intervalo de confianza de Wilson al 95 %**[10], que se comporta bien con muestras moderadas y con proporciones cercanas a cero —justo donde el intervalo «normal» de los cursos introductorios falla, como muestran Brown, Cai y DasGupta[11]. Las métricas son: exactitud (veredicto «exacta» sobre las juzgadas), **tasa de alucinación** (incorrecta más mal fundada), incompletas, abstenciones —distinguiendo las necesarias de las innecesarias—, fabricaciones en pruebas hostiles y citas fantasma. Y una regla de lectura: **«cero» nunca se afirma, se acota**. Cero fabricaciones en *n* pruebas hostiles no significa tasa cero: por la regla de tres, acota la tasa real a aproximadamente 3/*n* con 95 % de confianza. Un comité técnico refuta un «cero» en un minuto; un intervalo se sostiene.

Paso 6 — Comparación pareada McNemar antes de cambiar de modelo

Cuando cambia el modelo base, la estrategia de recuperación o el prompt del sistema, la pregunta correcta no es «¿subió el promedio?», sino «¿mejoró de verdad o cambió el ruido?». La prueba de McNemar[12] responde a eso: sobre las **mismas preguntas**, cuenta cuántas pasaron de fallar a acertar y cuántas de acertar a fallar, y entrega un valor *p* exacto sobre esos pares discordantes. Con motores no deterministas, es la única forma honesta de atribuir una mejora a un cambio. Nuestra política operativa es que ningún cambio de modelo entra en producción sin esta comparación contra la línea base; si degrada, se revierte.

Paso 7 — Congelado con fecha

Cada medición se publica con su fecha, la versión del motor, el tamaño del corpus al momento de medir y el hash del conjunto. **Nunca se retoca**: cuando el corpus crece o el motor cambia, una re-medición agrega una entrada nueva al historial y la anterior queda tal como se publicó. Es la única forma de que un comité que leyó una cifra en julio pueda volver en octubre y encontrarla idéntica, junto a la nueva.

Resultados públicos (medición de julio de 2026)

Las cifras siguientes corresponden a la medición del 2026-07-10 sobre un sistema en producción con corpus normativo chileno, y están congeladas con esa fecha[13]. Se reproducen aquí desde la tabla maestra de cifras del sitio; la tabla completa de veredictos con sus intervalos, la desagregación de las pruebas hostiles y la comparación pareada con la versión anterior están en la página del benchmark y en el paper técnico[14].

> DATO
>
> 278 consultas contra el sistema en producción, medición del 2026-07-10, sobre un corpus de 16 normas y 1.067 artículos al corte de la medición. Fuente: Innova y Cree, /empresas/benchmark/, consultado el 17 ago 2026.

MÉTRICA	RESULTADO (MEDICIÓN 2026-07-10)	CÓMO SE OBTUVO
Respuestas exactas	84,2 %	Juez independiente, rúbrica Stanford, con intervalo de Wilson al 95 % (en la página del benchmark)
Tasa de alucinación (incorrecta + mal fundada)	2,5 %	Ídem, con intervalo

MÉTRICA	RESULTADO (MEDICIÓN 2026-07-10)	CÓMO SE OBTUVO
Pruebas hostiles (artículos inexistentes, normas fuera de corpus, premisas falsas)	120 pruebas, 0 fabricaciones	Calificación por código; la lectura honesta no es «cero» sino una cota superior (≈ 3/*n*)
Mejora respecto de la versión anterior del motor	McNemar pareado, p = 0,0	Mismas preguntas, dos versiones; pares discordantes en la página del benchmark

resultados públicos de la medición de julio de 2026; toda cifra proviene de la tabla maestra de cifras y se publica congelada con su fecha.

Dos precisiones de lectura. La primera: la mejora entre versiones que reporta el valor *p* provino de correcciones estructurales de recuperación —no de un cambio de modelo— y la abstención innecesaria bajó a la vez sin que subiera la alucinación; ese es exactamente el patrón que la comparación pareada permite verificar. La segunda: los casos de alucinación de la corrida comparten un patrón —el sistema respondió desde un artículo vecino o una norma hermana y describió mal un detalle—, y en ningún caso inventó una norma o un artículo inexistente. Eso es lo que separa un error de matiz, detectable en segundos porque cada respuesta trae sus citas con enlace a la fuente oficial, de una fabricación.

La referencia internacional no es un ranking

Junto a nuestras cifras publicamos las del estudio de Stanford como orden de magnitud, no como comparación directa: aquellos sistemas operan sobre jurisprudencia estadounidense en lenguaje natural abierto; el nuestro, sobre normativa regulatoria chilena acotada. Lo comparable no es el número sino el **método**: misma definición de alucinación, mismo tratamiento de la abstención, misma exigencia de juez distinto del evaluado. El corpus, además, sigue creciendo —16

normas y 1.173 artículos al 2026-08-13— y por eso la próxima re-medición trimestral publicará su propio corte, sin tocar el de julio.

Qué NO mide este benchmark (limitaciones que declaramos)

Un benchmark que oculta sus límites es un argumento de venta. Estos son los nuestros, en el mismo orden en que un comité técnico los encontraría.

- **El juez es un modelo, no un abogado.** El acuerdo juez–humano (kappa) sobre esta corrida es el siguiente hito comprometido; hasta publicarlo, la tasa de alucinación es la que reporta un juez independiente calibrado por diseño pero no aún por medición.
- **Un dominio acotado.** El corpus evaluado es regulatorio y de tamaño conocido; cada dominio nuevo se mide con el mismo arnés antes de entrar en producción, y sus cifras son suyas, no heredadas.
- **Una corrida por versión.** Los motores no son deterministas: repetir la corrida da cifras parecidas, no idénticas. Por eso las cifras se leen con su intervalo y los cambios se validan de forma pareada.
- **No mide latencia, costo, seguridad ni fuga de datos.** Mide fiabilidad de respuestas con citas. La inyección de instrucciones, el aislamiento entre clientes o el manejo de datos personales tienen sus propias pruebas y sus propios artículos.
- **La detección de abstención es por frases.** Un motor que se abstiene con una redacción inusual puede quedar mal clasificado; el arnés permite reemplazar el listado y revisar a mano los casos dudosos.
- **El catálogo manda.** Las citas fantasma se recalculan contra el catálogo de artículos que se le entrega al arnés; si el catálogo está incompleto, la cifra hereda ese error.

Cómo puede un comprador exigir esto en un contrato

Lo que un contrato de servicios de IA suele traer es una cláusula de «mejores esfuerzos» sobre la calidad de las respuestas. Lo que debería traer es un **SLA de exactitud con método pactado**. Sus elementos mínimos:

ELEMENTO	QUÉ SE PACTA	POR QUÉ
Conjunto de evaluación	Tamaño, categorías (incluidas las hostiles), quién lo construye y quién lo custodia; el proveedor no lo ve antes de la primera corrida	Sin conjunto pactado la cifra no es reproducible
Definición de alucinación	Rúbrica escrita (incorrecta + mal fundada; abstención aparte)	Evita discutir después qué contaba como error
Umbral con intervalo	Tasa máxima de alucinación y de fabricación, expresadas con intervalo de Wilson y tamaño de muestra	Un umbral sin intervalo no es verificable
Versión fijada	Identificador de la versión del motor y del modelo base; aviso previo de cambio	Un alias flotante invalida la medición
Re-medición pareada	McNemar contra la línea base antes de cualquier cambio en producción; reversión si degrada	Es la única prueba de que el cambio no empeoró
Derecho a ejecutar	El comprador puede correr el arnés, en vivo, con sus propias preguntas	El proveedor que mide de verdad no teme la demostración
Congelado y publicación	Cada corrida se archiva con fecha, versión, hash y resultados; disponible en auditoría	Convierte la calidad en evidencia, no en promesa

elementos de un SLA de exactitud con método pactado para servicios de IA con citas.

Un detalle práctico: el conjunto de evaluación con normativa real **no se entrega al proveedor** por adelantado, y no se publica junto al motor que se evalúa. Se publican el esquema, el arnés y ejemplos sintéticos; el conjunto real lo custodia quien audita. Es la única forma de que la cifra mida al sistema y no su capacidad de estudiar para el examen.

Cómo reproducirlo con el arnés publicado

El arnés está publicado en GitHub bajo licencia MIT, en la versión 1.0.0, con cita en formato CITATION.cff y metadatos listos para Zenodo; el DOI está **en trámite** y se añadirá aquí y en el repositorio cuando se emita[15]. Es un solo archivo de Python, sin dependencias, con cinco subcomandos: `validar` (comprueba el esquema del conjunto), `correr` (ejecuta contra el endpoint, evalúa por código y llama al juez), `informe` (métricas con Wilson, informe Markdown y planilla para revisión humana), `comparar` (McNemar pareado entre dos corridas) y `kappa` (calibración juez–humano). Incluye un motor simulado sobre una norma ficticia y una prueba de punta a punta que corre en un minuto:

```
git clone https://github.com/Davidpintoi/benchmark-ia-verificable.git
cd benchmark-ia-verificable
bash pruebas/prueba-extremo-a-extremo.sh
```

Para correrlo contra un motor real se le indica el endpoint, el token por variable de entorno, la etiqueta de versión y —recomendado— el catálogo de artículos del corpus para recalcular las citas fantasma; las rutas dentro de la respuesta JSON del motor son configurables, así que sirve para cualquier proveedor que exponga respuesta, citas y confianza. Las fórmulas se pueden comprobar sin correr nada: la función de Wilson del arnés reproduce los intervalos publicados en la página del benchmark y la de McNemar reproduce su valor *p*. Quien quiera verlo funcionar contra nuestro sistema con las preguntas de su propio comité, incluidas trampas

sobre artículos que no existen, puede pedirlo: la demostración en vivo es parte de nuestro proceso comercial estándar.

> DATO
>
> 1.922 resoluciones judiciales en el mundo documentadas con contenido alucinado por IA (citas fabricadas u otros errores reconocidos por el tribunal). Fuente: D. Charlotin, AI Hallucination Cases Database, consultado el 17 ago 2026[^charlotin].

La lección de esa base de datos no es que la IA no sirva para el trabajo jurídico y regulatorio: es que ningún proveedor serio puede declarar el problema resuelto. Solo puede medirlo, publicarlo con su método y su fecha, y mejorarlo de forma demostrable. Ese es el estándar que aplicamos y el que un comprador tiene derecho a exigir.

1. NIST AI Risk Management Framework 1.0, función *Measure*. ↩
2. NIST AI 600-1, Generative AI Profile, riesgo «Confabulation». ↩
3. OWASP Top 10 for LLM Applications 2025, LLM09 Misinformation. ↩
4. Magesh et al., «Hallucination-Free?», JELS 2025; arXiv 2405.20362. ↩
5. Dahl et al., «Large Legal Fictions», J. Legal Analysis 2024; alucinaciones legales entre 58 % y 88 % según el modelo generalista. ↩
6. Google DeepMind, FACTS Grounding, 17 dic 2024. ↩
7. Jacovi et al., «The FACTS Grounding Leaderboard», paper (PDF): 1.719 ejemplos, 860 públicos; juicio en dos etapas (elegibilidad y fundamentación). ↩
8. Zheng et al., «Judging LLM-as-a-Judge», NeurIPS 2023. ↩

9. Landis y Koch, Biometrics 1977. ↩

10. Wilson, JASA 1927. ↩

11. Brown, Cai y DasGupta, Statistical Science 2001. ↩

12. McNemar, Psychometrika 1947. ↩

13. Innova y Cree, Benchmark anti-alucinación, /empresas/benchmark/. ↩

14. Innova y Cree, paper técnico del benchmark (PDF). ↩

15. benchmark-ia-verificable v1.0.0, GitHub; DOI Zenodo en trámite. ↩

16. Stanford RegLab, página del estudio. ↩

17. Charlotin, AI Hallucination Cases Database, corte 17 ago 2026. ↩

Puntos clave

01 «No alucina» no es una propiedad: es una medición con definición, conjunto de evaluación, versión del sistema, fecha e intervalo de confianza. Sin esas cinco cosas, la afirmación no es exigible.

02 El estado del arte (Stanford RegLab, FACTS Grounding) coincide en tres pilares: pruebas diseñadas para inducir el error, un juez distinto del sistema evaluado y una definición operativa que separa «incorrecto» de «mal fundado».

03 Nuestro protocolo evalúa por código todo lo que puede evaluarse por código (artículos inexistentes, materias fuera de corpus, premisas falsas, citas fantasma) y deja al juez solo lo que requiere criterio, con calibración contra revisión humana.

04 Toda tasa se publica con intervalo de Wilson al 95 %, «cero fallos» se acota en vez de afirmarse, y ningún cambio de modelo entra en producción sin comparación pareada McNemar contra la línea base.

05 El arnés completo está publicado bajo licencia MIT: cualquier comprador puede reproducir la medición contra el sistema de cualquier proveedor y pactar un SLA de

exactitud con método.

Preguntas frecuentes

¿Qué es una alucinación en un sistema de IA con citas?

¿Por qué no basta con que el proveedor diga que su IA «no alucina»?

¿Qué es el intervalo de Wilson y por qué importa?

¿Para qué sirve la prueba de McNemar en un benchmark de IA?

¿Puedo correr este benchmark contra el sistema de mi proveedor?

¿Qué debería exigir un contrato respecto de las alucinaciones?

Referencias

[1] Innova y Cree — Benchmark anti-alucinación: tasas medidas y publicadas con su metodología (medición 10 jul 2026). https://innovaycree.com/empresas/benchmark/. Consultado el 17 de agosto de 2026. PROPIO

[2] Innova y Cree — Paper técnico del benchmark (PDF). https://innovaycree.com/empresas/benchmark/paper-ia-verificable.pdf. Consultado el 17 de agosto de 2026. PROPIO

[3] benchmark-ia-verificable v1.0.0 — arnés de evaluación de alucinaciones para motores RAG por HTTP (GitHub, MIT; DOI Zenodo en trámite). https://github.com/Davidpintoi/benchmark-ia-verificable. Consultado el 17 de agosto de 2026. PROPIO

[4] Magesh, Surani, Dahl, Suzgun, Manning, Ho — «Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools», Stanford RegLab/HAI; 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

[5] Stanford RegLab — página del estudio «Hallucination-Free?». https://reglab.stanford.edu/publications/hallucination-free-assessing-the-reliability-of-leading-ai-legal-research-tools/. Consultado el 17 de agosto de 2026. PAPER

[6] 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

[7] 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

[8] Jacovi et al. — «The FACTS Grounding Leaderboard: Benchmarking LLMs' Ability to Ground Responses to Long-Form Input» (paper, PDF). https://storage.googleapis.com/deepmind-media/FACTS/FACTS_grounding_paper.pdf. Consultado el 17 de agosto de 2026. PAPER

[9] Zheng et al. — «Judging LLM-as-a-Judge with MT-Bench and [...] Arena», NeurIPS 2023 (arXiv 2306.05685): sesgos de posición, verbosidad y autopreferencia del juez. https://arxiv.org/abs/2306.05685. Consultado el 17 de agosto de 2026. PAPER

[10] NIST AI Risk Management Framework 1.0 (AI 100-1), 26 ene 2023 — función Measure. https://www.nist.gov/itl/ai-risk-management-framework. Consultado el 17 de agosto de 2026. NORMA

[11] NIST AI 600-1 — Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (jul 2024); riesgo «Confabulation». https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf. Consultado el 17 de agosto de 2026. NORMA

[12] OWASP Top 10 for LLM Applications 2025 — LLM09 Misinformation. https://genai.owasp.org/llm-top-10/. Consultado el 17 de agosto de 2026. NORMA

[13] Wilson, E. B. — «Probable Inference, the Law of Succession, and Statistical Inference», Journal of the American Statistical Association, vol. 22, 1927. https://www.jstor.org/stable/2276774. Consultado el 17 de agosto de 2026. PAPER

[14] Brown, Cai, DasGupta — «Interval Estimation for a Binomial Proportion», Statistical Science 16(2), 2001: recomienda Wilson frente al intervalo de Wald. https://doi.org/10.1214/ss/1009213286. Consultado el 17 de agosto de 2026. PAPER

[15] McNemar, Q. — «Note on the sampling error of the difference between correlated proportions or percentages», Psychometrika 12, 1947. https://link.springer.com/article/10.1007/BF02295996. Consultado el 17 de agosto de 2026. PAPER

[16] Landis, Koch — «The Measurement of Observer Agreement for Categorical Data», Biometrics 33(1), 1977: escala orientativa de kappa. https://www.jstor.org/stable/2529310. Consultado el 17 de agosto de 2026. PAPER

[17] Charlotin, D. — AI Hallucination Cases Database (resoluciones judiciales con contenido alucinado por IA). https://www.damiencharlotin.com/hallucinations/. Consultado el 17 de agosto de 2026. DATO

Citar este artículo

APA

```
Pinto, D. (2026, 17 de agosto). Cómo medimos las alucinaciones de una IA en producción (v1.0). Blog técnico de Innova y Cree. https://innovaycree.com/blog/ia-verificable/como-medimos-alucinaciones-ia-produccion.html
```

BIBTEX

```
@misc{pinto2026como,
  author       = {Pinto, David},
  title        = {Cómo medimos las alucinaciones de una IA en producción},
  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-medimos-alucinaciones-ia-produccion.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