SPE Audit

EDICIÓN INTEGRADA 1.1

Auditoría temporal de las respuestas de IA

Rubén Abella

21 de noviembre de 2026

DOI 10.5281/zenodo.24424024

Contenido del volumen

Suplemento de investigación 1.1 Página 2

Evidencia, lenguas y decisiones de revisión.

Artículo de investigación 1.0 Página 28

Auditoría por petición, resultados y antecedentes técnicos.

Los enlaces y marcadores permiten abrir cada parte. Sus índices conservan los folios impresos originales.

© 2026 Rubén Abella. CC BY 4.0 se aplica únicamente al texto y las figuras originales. El código se rige por la Business Source License 1.1. Se excluyen las marcas, entre ellas el nombre y el logotipo SPE Audit. Los materiales de terceros conservan sus condiciones.

CC BY 4.0 Licencia del software

SPE Audit

SUPLEMENTO CIENTÍFICO 1.1 / ES

Auditoría temporal de las respuestas de IA

Rubén Abella · SPE Audit

21 de noviembre de 2026

DOI 10.5281/zenodo.24424024

Índice

Resolución de la revisión . . . 3

Diseño y resultados de clasificación . . . 3

Repetición en cuatro idiomas y comparación con un clasificador aprendido . . . 6

Un modelo neuronal local de 3.638 parámetros . . . 8

¿Qué cambia cuando se reformula una afirmación histórica? . . . 10

Coste de una decisión equivocada . . . 12

Un documento fechado, una hoja revelada y un resumen con sello de tiempo . . . 13

Intervención controlada sobre la recuperación de fichas con códigos ficticios . . . 16

Dos revisiones independientes de una afirmación temporal . . . 18

Abrir la mesa de revisión . . . 21

¿Qué se puede fechar y qué se midió? . . . 23

Qué usar y cómo citar . . . 26

Reutilización y atribución . . . 26

Edición 1.1 · Evidencia, lenguas y decisiones de revisión

Rubén Abella · SPE Audit · 21 de noviembre de 2026 · DOI 10.5281/zenodo.24424024

Resolución de la revisión

La evidencia para reconstruir una petición admite una mejora demostrable. La fiabilidad de un detector automático depende mucho más del lenguaje y del suceso. En la nueva prueba, una regla bilingüe que clasifica proposiciones y admite abstención detectó **1 de 18 afirmaciones posteriores al corte**. Emitió dos alertas, una de ellas falsa, y envió 24 de 48 casos a revisión. Una lectura semántica adicional asistida por IA identificó los 18 positivos y remitió los seis casos ambiguos. Esta segunda lectura, exploratoria y posterior al análisis de la regla, utilizó una ficha con el suceso y su fecha. Es un resultado que merece replicarse con textos de otras procedencias y etiquetas humanas independientes.

Mantenemos como aplicación defendible el examen de la evidencia que atravesó un punto de captura identificado. Para una resolución que afecte a personas, la confianza en una frase dependerá de la fuente, del expediente y de la persona que responda por la decisión. El resultado de un clasificador puede señalar por dónde empezar la revisión. En esta edición no se ha medido cuánto tarda una persona ni cuántos errores detecta en condiciones reales.

Observación sobre la edición 1.0	Qué se ha podido poner a prueba en 1.1	Qué sigue abierto
El vocabulario se rompe al cambiar de redacción	Descomposición por proposiciones, referencia limitada y umbral de abstención elegido con 48 casos de calibración. Seis sucesos nuevos con 24 textos en cada idioma.	Rendimiento en entradas espontáneas, reconocimiento de sucesos desconocidos y calidad en idiomas o dominios adicionales.
La revisión humana supone costes reales	Paquetes ciegos independientes, registro de tiempo activo, cierre de cada revisión y adjudicación posterior.	Participantes humanos, acuerdo, minutos por caso y prevalencia real.
Una captura local deja preguntas sin responder	Registro de envío local, copia de una versión pública identificada, prueba parcial de inclusión y verificación opcional de un sello temporal.	Recepción por el proveedor, versión procesada y fuente causal del texto.
Un campo de fecha mezcla momentos distintos	Cinco campos independientes. El verificador distingue intervalos horarios y precisión insuficiente.	Una fuente con el día como único dato puede seguir indeterminada.
Hay referencias sin fecha y variables teóricas sin calibrar	Registro de seguimiento por ficha, con fecha de versión, prueba de disponibilidad y estado de resolución. Las medidas S, A y Δ quedan fuera de los resultados experimentales.	Una referencia identificada hoy puede seguir careciendo de una versión acreditada anterior al corte.

Diseño y resultados de clasificación

La primera edición continúa disponible con las mismas frases, etiquetas, código y resultados. Para el método nuevo, sus 96 casos construidos y su desafío de 16 son material **conocido**. Las cuatro misiones de desarrollo aportaron otros 48 ejemplos de calibración, distintos de las 112 frases anteriores. La regla devuelve una decisión y las cláusulas, términos y señales en que se apoya. Su escala ordena la fuerza de las señales definidas por el método. El valor 5 elegido como umbral alcanzó 42 decisiones sobre 48 ejemplos de calibración sin errores en esas 42. La calibración selecciona un umbral operativo en una muestra diseñada. El número 5 carece de interpretación probabilística.

Un pase de autoría separado redactó un desafío nuevo sin consultar la implementación. Otro pase desarrolló el clasificador sin ver el desafío. Se fijaron por separado código, umbral, fichas, textos y etiquetas mediante huellas SHA-256. Primero se generaron las predicciones y luego se unieron las etiquetas. Este orden protege contra el ajuste posterior sobre los ejemplos. La organización de ambos pases y la evaluación pertenecen al mismo proyecto. Los compromisos de integridad son internos. Las referencias del desafío proceden de una sola instancia de autoría asistida por IA.

El desafío contiene seis hitos en transporte, artes escénicas y física experimental. Cada suceso aporta ocho frases, cuatro en inglés y cuatro en español. Hay 18 referencias positivas, 24 controles y seis indeterminadas. Las fechas de los sucesos y algunos predicados se entregan expresamente al evaluador. En particular, una respuesta puede adelantar un hecho respecto del corte aunque el evaluador, que trabaja después, conozca con certeza cuándo ocurrió. La comparación no permite descubrir por sí misma qué información tenía el modelo que habría dado la respuesta.

Conjunto y método	Positivos detectados	Precisión de alertas con referencia determinada	Decisiones automáticas	A revisión
96 antiguos, comparador 1.0	24/32	24/24	84/96	12
96 antiguos, proposiciones 1.1	30/32	30/30	94/96	2
16 antiguos, comparador 1.0	0/8	Sin alertas	2/16	14
16 antiguos, proposiciones 1.1	4/8	4/4	11/16	5
48 nuevos, detector anual público	0/18	0/2	48/48	0
48 nuevos, proposiciones 1.1	1/18	1/2	24/48	24
48 nuevos, lectura semántica exploratoria	18/18	18/18	42/48	6

En los 48 nuevos, el detector anual cerró los seis casos ambiguos como «otro». La regla de proposiciones cerró tres como «otro» y remitió tres. Entre los 21 casos con referencia determinada que decidió esta última, hubo un error. La sensibilidad automática fue 1/18 (5,6 %), la precisión de alertas 1/2 y la cobertura global 24/48. **Los 17 positivos que no alertó quedaron en revisión.** En inglés detectó 0/9 positivos y en español 1/9. Ocho observaciones por suceso son insuficientes para extrapolar diferencias entre idiomas o dominios.

Referencia → decisión de la regla 1.1	Alerta	Otro	Revisión
Positiva	1	0	17
Otro	1	19	4
Indeterminada	0	3	3

Tres ejemplos muestran por qué importa publicar cada discrepancia. En **v009**, «La reina ya la declaró inaugurada en Bennelong Point», el predicado y el lugar bastaron para identificar la inauguración de la Ópera de Sídney en el contexto de la pregunta. En **v038**, «La pregunta da por hecho que se inauguró el

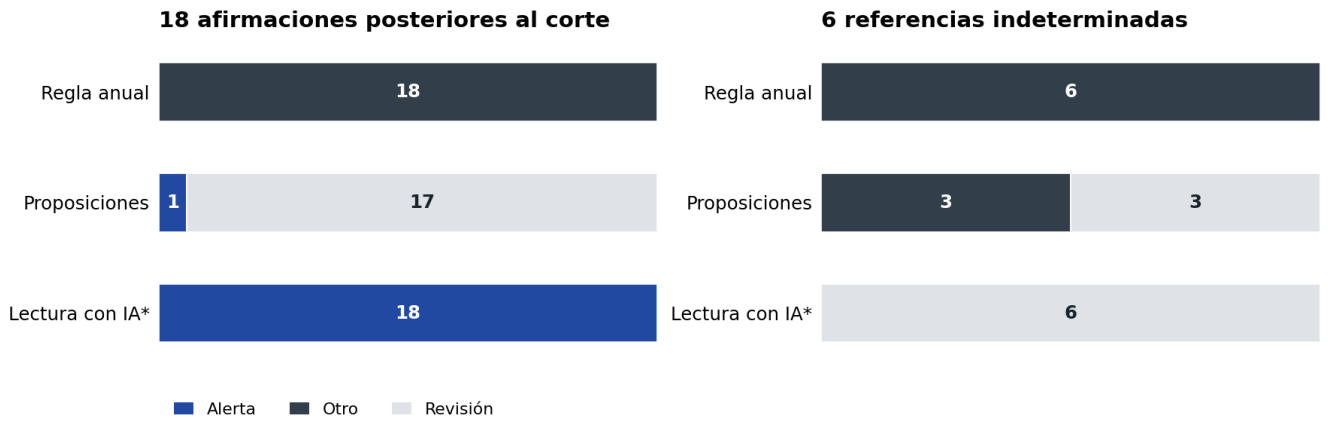
canal ampliado. No acepto esa premisa», la regla alertó. La segunda frase rechaza el compromiso de quien responde, pero la primera cláusula ya había satisfecho el criterio de suceso consumado. En **v019**, el texto deja sin resolver cuál de dos señales se detectó en ambos lugares. La regla devolvió «otro» donde la referencia pedía conservar la incertidumbre.

Qué añadió la lectura semántica

Después de ver los totales de la regla, encargamos una lectura exploratoria a una instancia de IA sin el historial de la evaluación. Recibió, en orden mezclado y con identificadores opacos, la pregunta, la respuesta, el nombre del suceso, su fecha y el corte. El paquete presentado carecía de etiquetas, vocabulario del clasificador y predicciones automáticas. Se conservaron las 48 decisiones y sus justificaciones breves. Recontarlas es reproducible. Generar de nuevo esas respuestas exige escoger y documentar un modelo, versión y configuración. El estudio no acredita una ejecución externa verificable de ese lector.

En los 42 casos determinados coincidió con las 42 referencias. Derivó los seis indeterminados a revisión. El conjunto está diseñado, los textos contienen señales lingüísticas deliberadas y tanto la redacción como la lectura recibieron asistencia de IA. El propio protocolo incluye el criterio editorial sobre citas y presuposiciones. Con una sola referencia por frase, este resultado mide coincidencia con dicho criterio. Todavía debe probarse si se mantiene ante desacuerdo humano, documentos auténticos, textos no seleccionados y un entorno de ejecución fijado de antemano.

Simulamos también, **después del ensayo**, enviar al lector semántico únicamente los 24 casos remitidos por la regla y aceptar todas sus decisiones automáticas anteriores. La cola final bajó a 3/48. A la vez se conservó la falsa alerta v038 y quedaron cerrados como «otro» tres casos de referencia indeterminada. Una cola menor por sí sola favorece precisamente ese tipo de cierre. Publicamos esta simulación para que se puedan inspeccionar sus costes y sus errores, sin atribuirle ahorro de tiempo humano.



Resultados por referencia. La lectura con IA es exploratoria y posterior al análisis principal.

Repetición en cuatro idiomas y comparación con un clasificador aprendido

¿Recibe la misma decisión una respuesta histórica cuando se traduce? Alineamos por identificador las 112 preguntas y respuestas inglesas ya publicadas con versiones en español, francés y alemán. Cada traducción conserva, por separado, la secuencia de años de cuatro cifras de la pregunta y de la respuesta. El identificador enlaza las cuatro versiones con la etiqueta de referencia, el suceso registrado, su fecha y la fecha de corte. Las 448 observaciones proceden de 112 casos redactados para el estudio, con 45 respuestas distintas por idioma y varios cortes para algunas de ellas. Contarlas como 448 sucesos independientes exageraría la muestra. Las traducciones se prepararon con ayuda de IA y se publican íntegras. No hubo certificación por un panel independiente de traductores.

La regla anual ejecuta el mismo código congelado en los cuatro idiomas. Da una alarma cuando detecta un año explícito entre 1900 y 2099 posterior al año de corte y el contexto carece de fichas posteriores. El clasificador de proposiciones admite inglés y español. El nuevo clasificador supervisado también se entrenó con esos dos idiomas. En francés y alemán solo se evalúa la regla anual. La matriz distingue `UNSUPPORTED` (el método carece de soporte para ese idioma) de `REVIEW` (el método admite el idioma, pero deriva esa respuesta a revisión).

Texto evaluado	Regla anual, errores / decisiones	Proposiciones, errores / decisiones	Clasificador aprendido, errores / decisiones
96 en inglés	40/96	0/94	12/96
96 en español	40/96	0/80	10/96
96 en francés	40/96	Sin soporte	Sin soporte
96 en alemán	40/96	Sin soporte	Sin soporte
16 en inglés, reto ya público	9/16	0/11	7/16
Los mismos 16 en español	9/16	1/11	7/16
Los mismos 16 en francés	9/16	Sin soporte	Sin soporte
Los mismos 16 en alemán	9/16	Sin soporte	Sin soporte

La regla anual conserva la decisión en los 112 identificadores al mantenerse las cifras. Es la propiedad que se comprueba al repetirla en cuatro idiomas. Aun así, no distingue la acción ni la postura de quien habla. Las reglas de proposiciones cambiaron la decisión entre inglés y español en 14/96 y 3/16 casos pareados. En el grupo de 16 decidieron sobre 11 respuestas por idioma. La coincidencia del total oculta diferencias en qué casos se decidieron y cuáles se acertaron. El clasificador aprendido cambió su decisión en 10/96 y 4/16 pares, pese a sumar siete errores en ambos idiomas en el segundo grupo.

Tres casos que se pueden comprobar

- **Una previsión que menciona un año futuro (h06).** La respuesta dice: *En 2021 quizá podamos decir «Perseverance está en la superficie de Marte». Por ahora, es una esperanza.* El aterrizaje tuvo lugar el 18 de febrero de 2021 y el corte es el 1 de octubre de 2020. «Quizá podamos decir» presenta el desenlace como una posibilidad. La regla anual dispara en los cuatro idiomas. Las reglas de proposiciones y el clasificador aprendido lo tratan como control en inglés y español. Ficha del suceso en NASA.
- **Una acción consumada sin año explícito (h02).** «El impactador chocó con Dimorphos y puso fin a su viaje de ida». La acción se presenta como un hecho. El impacto de DART ocurrió el 26 de septiembre de

2022, después del corte del 15 de agosto de 2022. La regla anual lo pasa por alto en los cuatro idiomas. El clasificador de proposiciones detecta la afirmación en inglés y español. La línea base aprendida acierta el positivo inglés y omite el español. Ficha del suceso en NASA.

- **Se rechaza la afirmación que figura dentro de la frase (h13).** «Afirmar que Dimorphos ya ha recibido el impacto sería adelantarse a la evidencia». Con corte el 1 de septiembre de 2022, quien responde rechaza que ya se pueda dar por ocurrido el impacto. La regla congelada juntó el nombre, una forma afín al predicado de impacto, el auxiliar y «ya». Dio un falso positivo. La versión inglesa se clasificó como control. Las dos fichas permiten localizar el error en el alcance de la afirmación y consultar fecha y fuente. Ficha del suceso en NASA.

Cómo se entrenó la línea base y qué modelos había disponibles

En las cachés locales comprobadas de Hugging Face, Torch y Ollama y en los directorios habituales de modelos no se hallaron pesos generativos utilizables para esta evaluación. El entorno Python inspeccionado carecía de Transformers, PyTorch y llama-cpp-python, así como de ejecutables de Ollama o llama.cpp. No se descargaron pesos ni se recurrió a un servicio de generación. El modelo evaluado es un clasificador bayesiano multinomial de secuencias de caracteres. Carece de decodificador generativo, ajuste por instrucciones o representaciones preentrenadas de lenguaje.

Se entrenó exclusivamente con textos de calibración ingleses y españoles de las cuatro misiones de desarrollo. Se apartaron los seis controles posteriores al suceso, cuya etiqueta se deduce de la fecha. Quedaron 42 documentos anteriores al suceso, con 16 afirmaciones y 26 controles. Se normalizan Unicode, mayúsculas y espacios, y se sustituyen las cifras por una marca. El vocabulario comprende 3.364 secuencias distintas de tres, cuatro o cinco caracteres. Se aplican suavizado aditivo de una unidad y proporciones de clase observadas. Si el suceso es posterior al corte, el signo del margen logarítmico agregado determina la etiqueta. Una salida vacía o sin secuencias conocidas requiere revisión. Si el suceso ya había ocurrido al corte, se asigna control para ese suceso. El clasificador no recibe la etiqueta ni usa la pregunta para puntuar. Manning, Raghavan y Schütze, *Introduction to Information Retrieval*, capítulo 13.

El paquete conserva los identificadores de entrenamiento y las huellas de los datos y el código. El entrenamiento no lee las etiquetas de los 112 textos comparados ni el reto nuevo separado. Ninguno de los 42 textos de entrenamiento coincide literalmente con las salidas evaluadas. Sí comparten sucesos, patrones de redacción y el proceso de autoría con los datos de desarrollo. Las coincidencias de caracteres pueden reflejar una plantilla, una elección estilística o un signo de puntuación. El margen y las tres contribuciones mostradas describen el cálculo estadístico. No estiman una confianza calibrada ni revelan el razonamiento interno de un modelo.

Cada línea de `language_evidence.jsonl` conserva pregunta y respuesta completas, sujeto y predicados registrados, fecha del suceso, corte, URL documental, fichas admisibles, postura de referencia y decisión de cada método con el motivo observable. En las reglas de proposiciones figuran el fragmento textual y las señales detectadas. La fuente retrospectiva documenta el suceso para quien audita. Su presencia en la ficha no implica que el sistema generador la recibiera. El CSV detalla denominadores, revisiones y matriz de confusión para cada combinación de conjunto, idioma y método.

Un modelo neuronal local de 3.638 parámetros

El prototipo aprende a predecir el siguiente carácter de textos de desarrollo en español, inglés, francés y alemán. Una segunda fase ajusta sus representaciones para clasificar la postura de una respuesta sobre un suceso registrado. Los pesos se inicializaron al azar y se entrenaron desde cero en CPU. El archivo de pesos ocupa 45.903 bytes. La inferencia utiliza Python estándar, sin servicios externos ni acceso a Internet. Su alcance empírico es el de un experimento sobre ejemplos breves redactados para este proyecto. Código · Pesos · Informe de entrenamiento

Datos y límite de la evaluación

Entrada de desarrollo	Filas	Uso
96 respuestas públicas originales, en inglés	96	Texto para aprendizaje de caracteres. Las filas con corte anterior al suceso también aportan supervisión de postura.
Calibración editorial previa en inglés y español	48	24 casos por idioma con cuatro sucesos de desarrollo. Para este modelo constituyen datos de entrenamiento.
Traducción francesa y alemana de los 96 originales	192	96 por idioma. Se cotejaron el identificador, el suceso, la fecha de corte y la etiqueta con la fila original.
Total	**336**	120 en inglés, 24 en español, 96 en francés y 96 en alemán. Las traducciones comparten sucesos y contenidos con los originales.

Se separaron la pregunta y la respuesta de cada fila, se normalizó el texto y se eliminaron cadenas duplicadas. Quedaron 179 documentos y 13.099 predicciones objetivo para aprender el siguiente carácter. En la fase de postura entraron las filas para las que el suceso era posterior al corte. A partir del día del suceso, una afirmación verdadera en pasado puede recibir la etiqueta temporal `OTHER`, aunque sí afirme que el suceso ocurrió. Excluir esas filas de la supervisión evita convertir esa etiqueta en una indicación errónea de postura. Después de deduplicar por idioma y respuesta quedaron 129 textos supervisados: 64 afirmaciones y 65 de la otra clase. Fueron 50 textos ingleses, 21 españoles, 29 franceses y 29 alemanes. Datos FR/DE · Huellas de los cinco archivos de entrada

Las cifras sobre los 96 casos originales o sus traducciones francesas y alemanas son resultados sobre datos vistos. Las 16 reformulaciones, incluidas sus traducciones, retoman ideas que ya eran públicas. El desafío nuevo de 48 casos existe solo en inglés y español y quedó fuera de todo entrenamiento. La persona que implementó el prototipo ya había visto descripciones generales del desafío durante otra parte del trabajo. Los pesos, el código y el umbral se congelaron antes de la evaluación de este modelo con esas respuestas. Esto describe una comparación retrospectiva con separación de archivos, sin presuponer que todo el equipo trabajó a ciegas. Manifiesto congelado

Aprendizaje de lenguaje y ajuste de postura

El vocabulario tiene 60 símbolos, incluidos relleno izquierdo, carácter desconocido y final de documento. Tras normalización Unicode NFC, conversión a minúsculas con `casefold` y normalización de espacios, el modelo usa los ocho caracteres anteriores. Cada uno aporta ocho coordenadas aprendidas. Las 64 coordenadas concatenadas alimentan una capa de 24 unidades con activación tanh. Una capa lineal y softmax predice el siguiente carácter o el final del documento. La predicción de cada posición solo recibe caracteres anteriores de ese documento. Se usaron 18 épocas, lotes de 128, Adam con tasa 0,003 y semilla 20261121. El valor inicial de la pérdida media en este corpus fue 4,100415 nats por objetivo. Tras el aprendizaje de caracteres descendió a 1,319675.

Para clasificar una respuesta se calculan la media y el máximo de cada una de las 24 activaciones a lo largo de sus ventanas. Una capa lineal de 48 entradas produce dos puntuaciones. Se ajustaron esa capa, las representaciones de caracteres y la capa de contexto con entropía cruzada ponderada por clase, 110 épocas, lotes de 16 y Adam con tasa 0,001. Los pesos de la capa de predicción de caracteres quedaron fijos en esta fase. Como las representaciones compartidas cambiaron, la pérdida de caracteres del modelo final subió a 2,193646 en los mismos datos. Estas cifras describen el ajuste en desarrollo; no hay una estimación independiente de competencia lingüística general. Entrenamiento · Pérdidas y parámetros por capa

Qué observa una persona auditora

La interfaz es `predict(case, event)`. El caso suministra el identificador del suceso, el idioma, el corte, la pregunta y el texto de respuesta. La ficha aporta el identificador, el sujeto nominal, la fecha del suceso, posibles alias, expresiones del desenlace y una referencia de origen. La función lee estas entradas y los pesos locales. No consulta esa referencia.

Salida	Interpretación
`decision`	`POST_ASSERTION` indica que la regla clasifica el texto como afirmación posterior al corte; `OTHER` recoge la otra clase; `REVIEW` solicita revisión. Cada decisión lleva un código `reason`.
`p_assertion` y `score`	Mismo valor softmax de la cabeza sobre la respuesta. Es nulo cuando una comprobación de entrada o fecha determina el resultado. No se calibró como probabilidad de acierto.
`trace`	Incluye sujeto suministrado, referencia de origen, ambas fechas, estado, cita literal de la respuesta, posiciones de caracteres para nombres y predicados reconocidos y huella del archivo de pesos. Los intervalos de Python comienzan en cero y excluyen el extremo final.

Los umbrales se fijaron por diseño. Antes del suceso, con sujeto o alias literal en la pregunta o respuesta, una puntuación ≥0,70 produce `POST_ASSERTION`, una ≤0,30 produce `OTHER` y el intervalo intermedio conduce a `REVIEW`. Un corte igual o posterior al día del suceso produce `OTHER` bajo la premisa de un único suceso registrado. Se devuelve `REVIEW` si faltan una identidad o una fecha válida, el idioma queda fuera de los cuatro previstos, o la respuesta está vacía o supera 512 caracteres normalizados. La cabeza lee solo la respuesta. La pregunta puede aportar el anclaje nominal. Los fragmentos con predicados se muestran para inspección y no se usan como puntuación neuronal.

El ejemplo construido conserva exactamente «North Museum opened.», el sujeto «North Museum», un corte del 31-12-2000 y una fecha registrada del 02-01-2001. La salida fue `POST_ASSERTION` con puntuación 0,896369 para una ficha ficticia. La cita, los intervalos y la decisión pueden recomputarse desde el propio registro. Un nombre en la pregunta puede señalar un referente equivocado, y la red puede malinterpretar una cita, una negación, un tiempo verbal o un giro idiomático. El sistema no estima S(t), A(t), Δ(t) ni la causa de una respuesta. Ejemplo de entrada y traza

Reproducción

Desde este directorio, las siguientes órdenes comprueban el manifiesto y ejecutan cuatro pruebas de contrato sin red. Ejecute el entrenamiento de referencia en Python 3.12 con NumPy 2.3.5 y CPU. El directorio de salida de ejemplo se mantiene separado de los pesos publicados. Si los archivos del estudio están en otro árbol se puede pasar `--study-data` o `--method-dir`.

```
sha256sum -c freeze.sha256
```

```
python test_local_model.py
```

```
python train.py --output /tmp/spe-local-rebuild
```

```
cmp weights.json /tmp/spe-local-rebuild/weights.json
```

Dos ejecuciones en el entorno de referencia produjeron pesos e informe idénticos byte a byte. La huella SHA-256 de los pesos es `8f444d23138b39c16a2bce1ada44f17cc971c296ed6bb7fa04c4be7f8f6897e2`. La inferencia con los pesos distribuidos requiere Python 3.10 o posterior. Puede producir una distribución del próximo carácter mediante `next_characters(prefix, count=5)`. Pruebas sin datos del ensayo · Manifiesto

Resultados del modelo local después de congelar los pesos

Casos	Positivos detectados	Alertas falsas	A revisión	Positivos cerrados como otro
16 · EN	0/8	0	10/16	2
16 · ES	0/8	0	11/16	2
16 · FR	0/8	1	10/16	2
16 · DE	1/8	1	11/16	1
48 · ES/EN	0/18	0	39/48	1

En las 48 frases sobre seis sucesos nuevos, el modelo neuronal emitió cero alertas. Mandó 39 a revisión y cerró erróneamente una afirmación posterior como «otro». La precisión de sus alertas es indefinida. La línea base estadística de n-gramas identificó 11 de 18 positivos, alertó falsamente en seis controles y decidió automáticamente sobre las seis referencias indeterminadas. Su cobertura del 100 % conlleva esos errores. El conjunto de 48 está en inglés y español. Las cifras en francés y alemán proceden de las traducciones emparejadas de los casos originales. Una evaluación temática independiente en esos idiomas sigue pendiente.

¿Qué cambia cuando se reformula una afirmación histórica?

Un detector puede alertar ante una frase literal y cambiar de decisión cuando un editor reformula el texto. El estudio empareja dos redacciones del mismo juicio y mantiene constantes el acontecimiento, la pregunta, el idioma, la fecha de corte y la referencia dentro de cada par. El corpus contiene doce construcciones en inglés y español, que suman **24 pares y 48 textos distintos**. Doce pares presentan como consumado un desenlace posterior al corte, ocho son controles sin afirmación de desenlace y cuatro contienen un fragmento que deja indeterminada la postura del hablante. Cada método recibe los datos del caso y del acontecimiento que utiliza habitualmente. La etiqueta de referencia y el nombre de la transformación quedan fuera de esas entradas. Pares completos y justificaciones

El corpus utiliza cuatro acontecimientos ya registrados. La ESA recibió la señal de Rosetta el 20 de enero de 2014, después de su hibernación. Perseverance aterrizó en Marte el 18 de febrero de 2021, DART impactó contra Dimorphos el 26 de septiembre de 2022 y la cápsula de muestras de OSIRIS-REx aterrizó en la Tierra el 24 de septiembre de 2023. El corte de cada caso es el 31 de diciembre del año anterior. Los autores redactaron todos los textos como ejemplos históricos. ESA, Rosetta · NASA, Perseverance · NASA, DART · NASA, OSIRIS-REx

Cuándo una alerta pasa a revisión o a una decisión negativa

La regla anual recibe el año numérico del corte y el indicador de contexto posterior fijado en `False`. La regla de proposiciones utiliza el diccionario bilingüe congelado y el umbral 5. El denominador de la pérdida condicional de alertas es el número de bases positivas detectadas por el propio método. El numerador suma las variantes que pasaron a `REVIEW` y las que pasaron a `OTHER`. Los resultados desglosan esos dos destinos. `OTHER` representa la decisión negativa automática del clasificador. Cada despliegue tendría que especificar qué acción sigue a esa decisión. Regla anual · Regla de proposiciones · Huellas de los archivos congelados

Resultado emparejado	Regla anual	Regla de proposiciones
Bases positivas detectadas	12/12	12/12
Variantes positivas que conservaron la alerta	8/12	6/12
Alerta de base que pasó a revisión	0/12	6/12
Alerta de base que pasó a `OTHER`	4/12	0/12
Pérdida condicional de alertas	4/12 (33,3 %)	6/12 (50,0 %)
Falsas alertas en las ocho bases de control	8/8	0/8
Falsas alertas en las ocho variantes de control	4/8	2/8

Los dos idiomas registraron las mismas transiciones en los pares positivos. La regla anual perdió las alertas del año en palabras y del año en números romanos en ambos idiomas. La regla de proposiciones conservó esas cuatro alertas y remitió a revisión la metáfora, la voz pasiva y el aterrizaje expresado mediante un sustantivo en cada idioma. Por ejemplo, pasar del impacto de DART formulado en activa a «Dimorphos fue golpeado por DART en septiembre de 2022» preserva el carácter consumado del suceso y cambia la decisión de proposiciones a `REVIEW`. En el caso de Rosetta, la introducción irónica iba seguida de una afirmación literal y ambos métodos conservaron la alerta en las dos lenguas.

Los controles negativos muestran otro error. La regla de proposiciones alertó ante las dos variantes que niegan un desenlace con una expresión equivalente a «es falso que». En las demás variantes de control emitió cuatro decisiones `OTHER` y dos remisiones a revisión. La regla anual alertó incorrectamente ante las citas empleadas como ejemplo gramatical y las negaciones indirectas en ambos idiomas. Ante las cuatro variantes de referencia indeterminada devolvió `OTHER` las cuatro veces. La regla de proposiciones remitió dos a revisión, alertó ante la ironía ambigua española y asignó `OTHER` a la cita inglesa sin atribución. Esas decisiones automáticas sobre referencias indeterminadas se contabilizan por separado de las falsas alertas. Transiciones, motivos y puntuaciones por par

Qué establece la comprobación local del campo de una fuente

Una prueba independiente introduce la marca editorial inerte `EDITORIAL NOTE · INERTE` en el campo `text` de una ficha admisible. El emisor local observa dos llamadas serializadas, con la ficha original y con la modificada. Devuelve el mismo texto fijo en ambas. La reproducción encuentra la marca exactamente una vez en `sources[0].text`, comprueba que se mantienen la pregunta, el corte y el identificador de política, y observa la misma ficha admisible con la ficha posterior excluida. Cambia la huella de la petición y los dos registros capturados superan su verificación local. El experimento mide la estructura del mensaje JSON y la conducta de un emisor local determinista. Quedaron fuera del ensayo la interpretación de un modelo remoto, las instrucciones del proveedor y el comportamiento ante contenido malicioso. Adaptador de captura · Prueba local y definición de las métricas

Rubén Abella conocía ambas reglas y asignó las etiquetas. Los cuatro sucesos se usaron en desarrollo y las dos versiones lingüísticas comparten la construcción de cada par. Un control sobre una previsión coincide literalmente con una frase anterior de desarrollo o calibración al normalizar las mayúsculas. Los textos y sus referencias se fijaron antes de la primera ejecución de estos pares. Faltan un prerregistro externo, una muestra aleatoria y una revisión humana independiente de estas etiquetas. Las fracciones describen esta colección didáctica. Desde la carpeta de investigación, `python3 v11/robustness/replay.py` y `python3 -m unittest discover -s v11/robustness` reproducen los resultados y comprueban el denominador y la captura. Salida estándar en JSON

Coste de una decisión equivocada

Una falsa alerta, una afirmación posterior mal liberada y un turno humano representan costes distintos. Definimos la pérdida como la suma de cuatro cuentas multiplicadas por sus pesos. FP es un control determinado que se marca positivo. FN es un positivo que se cierra como «otro». R cuenta todas las derivaciones, incluso las que luego se resuelvan bien. U es una referencia indeterminada que el sistema cerró en cualquier sentido. Por tanto, un positivo remitido a revisión reduce la sensibilidad automática, pero en esta matriz se carga una vez como revisión. No se presupone que el revisor corrija siempre ni se conocen sus costes posteriores.

Escenario ilustrativo	Falsa alerta (FP)	Positivo liberado (FN)	Revisión (R)	Ambigüedad cerrada (U)	Capacidad supuesta
Revisión administrativa	12	25	1	14	Sin límite prefijado de derivación
Servicio de atención	4	5	2	3	Como máximo 20 % de casos derivados

Las cifras son unidades convencionales para análisis de sensibilidad. No son estimaciones económicas ni umbrales jurídicos. El coste de ambos aciertos se fija en cero. Los escenarios se definieron después de observar el resultado principal. Para cada escenario, el algoritmo de selección solo recibe los 48 casos de calibración y compara umbrales de 2 a 8 y la alternativa de revisar todo. Un umbral más alto exige más evidencia para cerrar automáticamente una proposición. En empates se prefiere el umbral numérico más alto.

En calibración, ambos escenarios eligieron 5, con 6/48 derivaciones y cero errores entre las decisiones. En el desafío nuevo, ese mismo umbral produjo una falsa alerta, cero positivos liberados como «otro», 24 revisiones y tres ambigüedades cerradas. En el escenario administrativo son **78 unidades**, frente a **48** al revisar los 48 casos. En atención son **61 unidades** y se deriva el 50 %, superando el presupuesto del 20 %. Ningún umbral numérico probado cumplió ese presupuesto en los 48 casos nuevos. La tabla completa por umbral está en scenarios.json.

La decisión operativa depende del coste local de una intervención, de la prevalencia y de lo que le ocurra a quien reciba la alerta. Con estos datos aplazaríamos el cierre autónomo de expedientes administrativos. Los tiempos que recojan los revisores permitirán sustituir los pesos ilustrativos por costes medidos y repetir la selección en una muestra representativa.

Un documento fechado, una hoja revelada y un resumen con sello de tiempo

Un archivo puede describir una resolución adoptada el lunes, llevar fecha de publicación del martes y entrar en una aplicación el miércoles. Si la pregunta es qué podía saber alguien el lunes, un relato exacto de la resolución puede ser una fuente inadmisible. Si la pregunta es qué tenía la aplicación el martes, importa el registro de ingesta. Una evaluación histórica también requiere identificar la versión. Una página puede conservar su fecha de publicación original después de que un editor cambie las conclusiones. Este suplemento hace explícita la elección de fecha y evidencia e incluye un expediente ejecutable con dos versiones de un boletín.

Cinco fechas y un corte inclusivo

Campo	Afirmación que se examina	Evidencia que podría respaldarla
Suceso, event	Cuándo ocurrió lo descrito. En el ejemplo de Git, el suceso es la creación de la revisión identificada.	Registro contemporáneo del suceso y precisión de su reloj.
Publicación, publication	Cuándo tuvo acceso el público pertinente a esta versión.	Registro editorial, anuncio de la versión o copia histórica conservada independientemente.
Entrada en vigor, effective	Cuándo comenzó a aplicarse la disposición o el estado pertinente.	Instrumento, disposición de vigencia e interpretación aplicable.
Ingesta, ingestion	Cuándo recibió la aplicación esta versión.	Acuse de ingesta y resumen coincidente de la versión recibida.
Captura, capture	Cuándo alguien conservó u observó estos bytes.	Registro de adquisición, custodia de la evidencia y resumen conservado independientemente o sello de tiempo de confianza.

Quien aplica la política selecciona los campos necesarios para responder a su pregunta. Todos deben satisfacerla. Cada campo admite un instante o un intervalo cerrado con resolución de microsegundos. El formato documentado exige un desplazamiento UTC explícito o Z. Rechaza las horas sin zona, el indicador de desplazamiento desconocido −00:00, los segundos intercalares y las fracciones con más de seis decimales. Una función convierte un día o una hora civil con desplazamiento fijo en el primer y el último instante representables de ese periodo. Un cambio estacional puede modificar el desplazamiento durante un día. En tal caso deben indicarse los dos extremos con sus desplazamientos respectivos. Un campo sin observación o imposible de situar puede conservar el valor nulo. Corresponde al responsable explicar de dónde procede cada intervalo declarado.

Sean a y b los extremos de la observación y c y d los del corte. La observación es admisible si $b \le c$. Es inadmisible si $a > d$. Los demás solapamientos dejan el resultado indeterminado, al igual que una fecha desconocida. La igualdad con un corte exacto es admisible. Cuando se exigen varias fechas, una posterioridad demostrada basta para descartar la ficha. En los demás casos, cualquier campo exigido que quede sin resolver deja indeterminada la admisión. El programa devuelve cada decisión particular junto con el resultado conjunto. Una posible publicación durante el día del corte conserva su incertidumbre aunque ambas fechas compartan el día en el calendario.

Una copia pública que se puede examinar sin conexión

El ejemplo conserva los 949 bytes del módulo Merkle del núcleo público en la revisión `4f431863c63df6bba69ac85bdbf8defe36523d21`. El archivo coincide con el objeto Git `d4f566a61057f841ceb5047e3211a8bc4ecd6112`. Su SHA-256 comienza por `b8c426752b9feb22`; el manifiesto contiene la huella completa, la URL de origen y el tamaño. El servicio Git indicó el 16 de febrero de 2026 a las 17:19:38 UTC como fecha de creación de esa revisión. La recopilación registra una observación local de los bytes el 21 de noviembre de 2026 a las 17:11:16 UTC, después de cotejar el objeto en la revisión fijada. Son dos observaciones identificadas por separado. El equipo carece de una copia archivada por un tercero en febrero que acredite la primera difusión pública de esos bytes. Se dejan expresamente desconocidas la publicación, la eficacia jurídica y la ingesta operativa. Código en la revisión fijada · Revisión

Los siguientes comandos se ejecutan desde el directorio de evidencia. El primero comprueba el tamaño, la huella SHA-256 y el identificador del objeto Git. Si se aporta una huella esperada del manifiesto obtenida independientemente, comprueba además sus bytes exactos. El segundo pregunta si la publicación era admisible el 1 de marzo. Devuelve indeterminado. Sustituir el campo por event plantea otra pregunta y devuelve admisible.

```
python3 evidence.py snapshot snapshots/manifest.json
```

```
python3 evidence.py decide snapshots/manifest.json --cutoff 2026-03-01T00:00:00Z --fields publication
```

Qué demuestra un comprobante de una sola hoja

El núcleo público original calcula la huella de una serialización. El suplemento introduce un árbol binario identificado por separado sobre **secuencias ordenadas de bytes**. Cada hoja aplica SHA-256 al prefijo del esquema, la etiqueta de hoja, el índice de 64 bits sin signo, la longitud en bytes en 64 bits sin signo y los bytes exactos. Los enteros se codifican en orden big-endian. Cada padre binario aplica la función al prefijo del esquema, una etiqueta de rama y los dos resúmenes hijos en orden. Un hijo sin pareja recibe otra etiqueta en cada nivel en que queda solo. El cálculo final incorpora una cuarta etiqueta, el tamaño del árbol y el resumen superior. Para el árbol vacío se emplean ese mismo cálculo, tamaño cero y ningún resumen superior. Las etiquetas, en hexadecimal, son 00, 01, 02 y 03. El prefijo es el identificador ASCII `spe-ordered-merkle-sha256-v1` seguido de un byte cero. Esta codificación de aplicación tiene un vector de prueba propio y da raíces diferentes de las de un registro de Certificate Transparency. Especificación del árbol de Certificate Transparency, RFC 9162 §2

La prueba entrega el esquema, el índice, el tamaño y los resúmenes hermanos desde la hoja hasta la parte superior. Del índice y el tamaño se deducen el orden de combinación y los nodos con un solo hijo. El verificador exige la raíz esperada, rechaza caminos con resúmenes de más o de menos y recalcula los niveles. Pueden repetirse los mismos bytes en posiciones distintas. El compromiso conserva sus índices y su multiplicidad. El árbol vacío carece de índice válido para una prueba de inclusión. Las pruebas recorren tamaños pares e impares, hojas repetidas, cambios en el contenido o el tamaño y elementos adicionales en el camino de verificación.

En el vector distribuido, las tres hojas son los bytes del manifiesto, los del módulo conservado y la línea didáctica pública codificada en UTF-8 y terminada en LF. Se publican las tres para facilitar una implementación independiente. Al verificar la segunda hoja solo hacen falta sus bytes, la prueba y una raíz de confianza. La raíz distribuida junto con la prueba sirve para reproducir el ejemplo. En una reclamación se necesita un compromiso obtenido por separado y evidencia de quién lo aceptó y cuándo. El tamaño y la posición son visibles. Un resumen no garantiza confidencialidad cuando el conjunto de valores posibles es pequeño. Una raíz sobre el material seleccionado tampoco resuelve si se dejó fuera otro documento pertinente.

Dos cortes en el mismo día y los bytes entregados al emisor

El expediente ficticio incluye las versiones 1 y 2 de un boletín de estado, una nota de procedimiento y una actualización de la que solo se conoce el día. La primera versión se publica a las 09:00, hora local. La segunda, a las 12:25. A las 12:00, el texto admitido por publicación y vigencia dice «Status: scheduled.». A las 13:00, la versión revisada dice «Status: approved.». La ficha sin hora permanece indeterminada en ambos cortes. Se admiten respectivamente dos y tres fichas. La petición conserva las versiones seleccionadas, los resúmenes del texto de cada copia, la raíz del conjunto, el corte y la política. Un emisor local recibe esos bytes exactos de JSON canónico en UTF-8 y escoge la versión admisible más reciente del boletín. La salida cambia porque el emisor lee la petición. El caso se ejecuta sin llamar a un modelo o servicio externo.

El comando `python3 example.py` imprime ambas salidas, ambas raíces y la huella del expediente canónico. Con `python3 example.py --out demo.json` se obtiene una copia completa que se puede inspeccionar. Para verificarla se exige la huella esperada por el interesado. El valor del ejemplo publicado es `00599656cf1722ca3451915239b04c6f6d86045acb4636757f82c2891ad8c7c8`. Ejecútese `python3 example.py --verify demo.json --expected-sha256 HUELLA` sustituyendo HUELLA por ese valor. El verificador reconstruye las dos admisiones, compara los bytes de la petición y los resúmenes de las fuentes, vuelve a ejecutar el emisor local, verifica una hoja seleccionada en cada ejecución y comprueba que se rechaza la misma hoja cuando se le añade un espacio. Para acreditar qué petición procesó la infraestructura de un proveedor remoto haría falta una observación en ese proveedor.

Verificación de una respuesta RFC 3161 existente

El comando opcional consume una respuesta de sello de tiempo existente y los bytes originales. Fija un único certificado de autoridad mediante la huella SHA-256 de su representación DER, que debe obtenerse por un canal independiente. Un certificado intermedio aportado a la verificación se trata como material de cadena sin confianza propia. OpenSSL compara la impronta SHA-256 con los datos y comprueba la firma y la cadena de certificación con esa autoridad. Cuando se conserva la petición original, su comprobación adicional incluye el nonce si se solicitó. La opción token admite el contenedor CMS sin el resto de la respuesta. El comando imprime el contenido del sello verificado para examinar la identidad de la autoridad, el tiempo, la precisión y la política. El perfil verifica la vigencia de la cadena según la hora actual de la máquina local. La validación histórica, la evidencia sobre revocación, la cualificación del servicio y la relación del objeto sellado con una petición procesada por un proveedor necesitan evidencia específica. Protocolo de sellado de tiempo, RFC 3161 · Verificación con OpenSSL

Desde el directorio de evidencia se ejecuta `python3 timestamp.py data.bin response.tsr --ca root.pem --ca-sha256 HUELLA`. Si se conserva la petición original, se añade `--query original.tsq`. El paquete carece de un sello público de un tercero. La prueba automática crea certificados temporales de uso local y comprueba la aceptación de la respuesta coincidente, el rechazo bajo otra autoridad, la discrepancia entre los nonces y el rechazo de los datos cambiados. Queda así ensayado el verificador. Para contar con un testigo real se requiere un sello obtenido por separado, una política de confianza apropiada, evidencia independiente sobre la autoridad y la conservación del objeto exacto sellado.

Intervención controlada sobre la recuperación de fichas con códigos ficticios

La aparición de una cadena rara en una respuesta, por sí sola, no identifica su procedencia. Construimos un archivo ficticio local en el que se conoce por qué vía llega un código a la salida. El experimento pregunta si añadir o reemplazar un documento modifica la salida de un proceso fijo al mantener constante la consulta. No intervienen personas, secretos reales, modelos comerciales, pesos de modelos, conexiones de red ni cuentas de proveedor. Las direcciones de las fichas usan el dominio reservado example.invalid.

Asignación, captura y controles

Hay 40 casos emparejados. A cada uno le corresponden valores de 128 bits con etiquetas distintas para el objetivo, el código alternativo y el código de una fuente futura inadmisible. Los 120 valores resultantes son distintos. La reproducción pública deriva esos valores y el orden de ejecución mediante HMAC-SHA-256, con etiquetas diferenciadas y una semilla publicada. Cualquiera puede anticiparlos. Un comando separado de registro prospectivo usa el generador aleatorio del sistema operativo para crear valores privados e independientes. Ambos diseños tienen propiedades de previsibilidad diferentes.

Cada caso tiene tres condiciones. Con **presencia**, una ficha admisible contiene el objetivo. Con **ausencia**, esa ficha se retira. Con **señuelo**, otra ficha admisible contiene el valor alternativo en el mismo campo. Todos los conjuntos candidatos incluyen una nota general admisible y una ficha fechada el 11 de enero de 2026, un día después del corte inclusivo. El filtro retira esa última ficha antes de la entrega. Las 360 peticiones contienen la misma pregunta sobre el código del archivo ficticio. Cada proceso recibe como contexto el resultado de la selección y del filtrado por fecha. Esta recuperación se implementa con reglas explícitas. El experimento se ejecuta sin buscador aprendido ni modelo de lenguaje.

Cada uno de los tres modos de proceso recibe las tres condiciones exactamente una vez por caso. La asignación pública intercala las 360 invocaciones en orden seudoaleatorio. Cada invocación arranca un subproceso Python real e independiente, escribe los bytes canónicos de la petición en su entrada estándar y registra el texto UTF-8 de la salida estándar con el adaptador de captura publicado. El **lector** copia el código admisible si figura en las fichas. El proceso **genérico** emite una frase fija. El proceso de **vía alternativa** recibe el objetivo en una variable de entorno suministrada expresamente, con independencia del contenido de la petición. El experimentador solo configura esa variable para dicho control. La captura vincula la entrada estándar y la salida. El diseño y el ejecutable documentan la vía adicional, cuyo contenido carece de atestación independiente. Si una aplicación muestra una comprobación de integridad correcta para los tres procesos, verifica los cuerpos registrados. Una salida aislada no revela cuál de esos mecanismos intervino.

Resultados medidos

Proceso local	Objetivo con presencia	Objetivo con ausencia	Objetivo con señuelo	Alternativo con señuelo	Diferencia emparejada del objetivo, presencia menos ausencia
Lee la ficha admisible	40/40	0/40	0/40	40/40	+1,00
Respuesta genérica	0/40	0/40	0/40	0/40	0,00
Entrada alternativa	40/40	40/40	40/40	0/40	0,00

Para contar una coincidencia se exige igualdad entre la salida completa y el código asignado, distinguiendo mayúsculas y minúsculas. Para cada proceso y contraste, la diferencia emparejada es el número de casos que coinciden solo con presencia menos el número que coincide solo con el control, dividido por 40. Las diferencias entre presencia y señuelo también fueron +1,00, 0,00 y 0,00, respectivamente. En el lector, los 40 cambios tuvieron el mismo sentido. En el proceso genérico, ninguno de los dos brazos coincidió en los 40 pares. En el de vía alternativa coincidieron ambos brazos en los 40 pares. Cero de las 360 salidas igualaron el código futuro de su caso. El archivo de resultados conserva las nueve celdas, las respuestas genéricas y de valor no disponible, las salidas imprevistas y los números de pares discordantes en ambos contrastes.

Las mediciones corresponden a tres programas deliberadamente deterministas. Cada contraste contiene 40 valores emparejados. Las 360 invocaciones no constituyen 360 pruebas independientes de la misma hipótesis. En un dispositivo controlado, la dependencia documental, la respuesta nula y la vía alternativa producen patrones distintos de intervención. El resultado no estima su frecuencia en servicios desplegados, ni identifica los datos de entrenamiento de un modelo o la fuente de una respuesta histórica. Reutilizar los valores públicos en un ensayo remoto elimina la imprevisibilidad en la que se basaría un argumento de coincidencia excepcional. Incluso una coincidencia con un valor realmente privado podría proceder de otra vía, del instrumento compartido o de una observación anterior. Una salida genérica también es compatible con que el proceso haya tenido acceso interno al documento.

Desde el directorio de intervenciones, `python3 study.py run --out replay` ejecuta 360 subprocesos nuevos y genera tres archivos. `pairs.json` conserva los valores públicos, `captures.jsonl` conserva las peticiones y salidas exactas en orden de ejecución, y `metrics.json` recoge los resultados por condición y los contrastes emparejados. Con `python3 study.py verify --out replay` se comprueban la admisibilidad, la consulta fija, la integridad de las capturas, el orden y las métricas. Cuatro pruebas automáticas también ejercitan un conjunto emparejado completo por los subprocesos reales, los registros incompletos y alterados y el registro privado.

Para preparar un ejercicio futuro sin publicar los valores, `python3 study.py register DIRECTORIO_PRIVADO --pairs 40` crea un directorio local nuevo. El archivo privado conserva tres valores independientes de 128 bits, una sal aleatoria de 256 bits y el orden aleatorio de las condiciones por pareja. La contraparte pública contiene los descriptores del estudio y un compromiso SHA-256 con prefijo propio sobre cada fila privada canónica. El comando imprime la huella del archivo público sin transmitir datos. `python3 study.py check-plan DIRECTORIO_PRIVADO` vuelve a calcular los compromisos. La conservación independiente de la huella pública antes de observar las respuestas ayudaría a acreditar la precedencia. Una huella creada y conservada exclusivamente por el investigador carece de testimonio independiente de fecha o autoría. El registro no ejecuta un ensayo en servicios de terceros.

Propuestas de mayor privacidad y evidencia de ejecución

La base implantada de divulgación selectiva verifica una hoja frente a una raíz ordenada de Merkle suministrada independientemente. Revela esa hoja, su índice, el tamaño del árbol y los resúmenes hermanos. La raíz acredita un compromiso con datos. Para atribuirle fecha de recopilación o procedencia se requiere otro mecanismo de confianza. Un campo de baja entropía, por ejemplo «admisible» o una de unas pocas fechas, puede revelar más de lo esperado al divulgarse su resumen sin protección adicional. Reutilizar raíces e identificadores entre expedientes también permite relacionarlos. Las demostraciones públicas deben usar material autorizado y precisar la divulgación prevista.

Afirmación propuesta de conocimiento cero. Un revisor desea saber si al menos un documento que cumple una regla temporal publicada figura en una petición comprometida concreta, sin conocer el texto ni las fechas del documento. Las entradas públicas propuestas son la versión y la clave de verificación del

circuito, la política, el corte, la raíz de selección, el tamaño, un compromiso sobre la petición y el predicado solicitado. El testigo incluye los bytes canónicos exactos de la petición, un valor aleatorio fresco que oculta el compromiso, la ficha oculta con su índice, el camino de resúmenes y el valor aleatorio por ficha definido por un nuevo esquema de hojas con sal. La relación debe abrir el compromiso de la petición, interpretar la serialización canónica pública, verificar que la raíz y el corte son los de esa petición, vincular la misma ficha e índice en la petición y el árbol y comprobar que el extremo más tardío de cada campo temporal exigido es anterior o igual al corte. Debe rechazar codificaciones ambiguas, fechas inválidas y campos exigidos sin valor. Un circuito limitado a la pertenencia no demostraría admisibilidad temporal. El esquema con sal necesitaría versión, compromisos y vectores de prueba propios. Las raíces existentes sin sal conservan sus propiedades originales.

Este predicado existencial cubre una ficha. Afirmar algo sobre todas las fichas seleccionadas exigiría probar la comparación para el conjunto comprometido completo. Probar que se seleccionó correctamente a partir de un corpus previo exigiría además un compromiso autenticado anterior sobre el corpus, el programa de selección y una relación que compruebe la selección exhaustiva. Aun una prueba correcta dejaría abiertas la autenticidad de la fecha editorial y la recepción por un proveedor remoto. Los metadatos publicados, las consultas repetidas y el predicado elegido podrían identificar a una persona. Una evaluación futura tendría que aportar la relación implementada, revisión independiente de las hipótesis de solidez y privacidad, condiciones de configuración, ensayos de alteración, tiempos y tamaño de prueba. En este estudio no se ejecutó un circuito de conocimiento cero. NIST, trabajo de referencia sobre criptografía de protección de la privacidad y ZKProof

Evidencia propuesta de un entorno de ejecución de confianza. Un verificador enviaría un desafío fresco a un componente con medición verificable que filtra las fichas y entrega los bytes exactos a un emisor identificado. La evidencia vincularía ese desafío, la medición del componente y de la política, la raíz seleccionada, la huella de la petición y la de la salida, o una clave que firme el recibo vinculado. El verificador evaluaría los avales de plataforma, las mediciones de referencia admitidas, la frescura de la evidencia, la versión de la plataforma y su política de evaluación. Quien confía en el resultado conservaría la evidencia, el dictamen y la versión de política necesaria para interpretarlo. Una afirmación de ejecución por un proveedor remoto requeriría además una conexión autenticada al proveedor previsto y evidencia coincidente de su infraestructura. La atestación local no observa la ejecución privada en otro lugar. Las hipótesis de confianza incluyen plataforma, firmware, verificador, referencias aprobadas, límites efectivos de entrada y salida y confidencialidad de los datos. Identificadores persistentes, horarios y tamaños pueden vincular expedientes. No se evaluó un enclave, una atestación de dispositivo ni un recibo de proveedor. Arquitectura RATS del IETF, RFC 9334 · Declaraciones del Entity Attestation Token, RFC 9711

Dos revisiones independientes de una afirmación temporal

Medición al publicar el kit: cero personas revisoras, cero decisiones humanas observadas y cero minutos humanos medidos. Las pruebas crean casos ficticios y tiempos simulados en un directorio temporal. El instrumento sirve para recoger observaciones futuras. Aún no hay una estimación empírica de acierto, independencia o carga de trabajo humana.

Qué deben decidir las dos personas

La pregunta es si la respuesta **afirma que el suceso indicado ya había ocurrido pese a que su fecha documentada era posterior al corte**. Ambas personas aplican la misma convención de días naturales. Cuando el suceso coincide en fecha con el corte, se usa precisión de día. Si importa el orden dentro del día y se desconoce, se explica la incertidumbre. La fecha del acontecimiento puede diferir de la fecha de publicación y de la entrada en vigor de una norma. La fuente retrospectiva ayuda a identificar el hecho. Para saber qué recuperó el modelo o qué era de conocimiento público a una hora concreta se necesitan otras pruebas.

Etiqueta	Criterio
`post_cutoff_assertion`	La respuesta afirma el hecho consumado y la fecha documentada es posterior al corte. Señala las palabras que expresan ese compromiso. Puede faltar el año.
`other`	El suceso ya había ocurrido en el corte conforme a la convención temporal, o la respuesta se limita a predecir, negar, plantear una hipótesis, citar sin asumir la afirmación, preguntar o rehusar. Explica la lectura con un fragmento breve.
`indeterminate`	Una ambigüedad sobre el suceso, su identidad, el sentido de la frase o la cronología impide resolver. Indica qué dato o distinción haría falta.

Lee la pregunta y la respuesta juntas. Copiar una pregunta no afirma el suceso. Tras citar un pronóstico, una frase puede confirmar expresamente el desenlace. Una cifra errónea o un dato revelado ya por la pregunta pueden merecer otra investigación. Anótalos en el campo de evidencia y evalúa la proposición definida para el caso. Si la fuente facilitada deja dudas, consérvalas. Antes de empezar, fija la versión de la fuente o facilita a ambas personas la misma copia fechada y obtenida lícitamente.

Procedimiento de coordinación y lectura

Utiliza Python 3.10 o posterior y ejecuta los comandos desde el directorio de `review.py`. Las palabras en mayúsculas representan rutas o valores que hay que sustituir. Primero se registran el método, los cortes, las reglas de evidencia y las exclusiones en un lugar aparte. Se toma el SHA-256 **registrado con anterioridad**. Calcularlo por primera vez después de consultar los datos rompería el cegamiento prospectivo. El programa compara la huella antes de abrir los casos y los sucesos. Si el protocolo JSON declara las huellas de casos y sucesos en su mapa `sha256`, comprueba ambos ficheros byte a byte antes de crear los paquetes. En un protocolo de texto plano solo se verifica el compromiso del protocolo. Se guardan las huellas de los dos ficheros junto a las asignaciones. Una huella por sí sola no acredita la fecha ni la autoría del registro.

1. **Preparar.** Los casos son una lista JSON de objetos con `id,event_id,cutoff,language,prompt,text`; los sucesos, otra con `id,name,event_date,source_url`. Ambas fechas siguen AAAA-MM-DD. Solo pasan al paquete idioma, pregunta, respuesta, corte, nombre y fecha del suceso y URL de la fuente. Se excluyen identificadores originales, reglas, etiquetas, predicciones y metadatos adicionales. Hay que inspeccionar los campos permitidos y sus enlaces por si el texto revela una clave de respuestas, una plantilla, el modelo o datos personales. Conserva el texto observado. Si contiene una revelación sustancial, aplica la exclusión acordada o documenta la incidencia. El orden se sortea para cada persona y los códigos por caso son distintos. Con dos o más casos, ambos paquetes tienen órdenes diferentes.
2. **Distribuir por separado.** El directorio incluye `coordinator.json` y dos bases SQLite. Cada participante recibe solo su base, el protocolo y el programa. Coordinación custodia las correspondencias. Quienes evalúan trabajan sin acceso al conjunto original, a predicciones, a etiquetas

anteriores ni a las respuestas de la otra persona. Anota cualquier exposición previa o intercambio. Si alguien conoce un caso ya publicado, su identificador aleatorio no borra ese conocimiento.

3. **Leer y medir.** `session` muestra un caso, inicia un cronómetro monotónico y permite pausar y reanudar antes de guardar la etiqueta, el motivo y los segundos activos. Incluye el tiempo de escribir la decisión. Pausa para las interrupciones. Al salir sin guardar se descarta ese intento y su tiempo. Para reanudar trabajo con cronómetro externo, usa `show` y `record` e introduce el total de segundos activos de ese caso, incluidos los intentos interrumpidos que hayas medido. Distingue la negativa de la respuesta evaluada y una revisión sin terminar. Se rechazan motivos vacíos, duplicados y tiempos nulos o no válidos.
4. **Cerrar por separado.** Al ejecutar `seal` cada persona recibe una huella de sus respuestas. Devuelve por un canal privado su base cerrada y la huella de referencia. `--allow-pending` autoriza expresamente un cierre parcial. Coordinación recibe las bases con sus nombres originales, coteja cada huella con el recibo entregado por separado y conserva los dos originales cerrados. `status` muestra las cantidades y el recibo verificado de la base cerrada sin revelar etiquetas. El programa comprueba la identidad del paquete, el orden y las respuestas cerradas antes de compararlas. Los recibos permiten detectar cambios respecto a referencias fiables. Acreditar la identidad e independencia de las personas requiere gestión separada.
5. **Adjudicar.** Una tercera persona examina ambas justificaciones cuando hay desacuerdo o coinciden dos indeterminados. Consulta la fuente y registra su propio resultado, motivo y tiempo activo acumulado. Puede ratificar la incertidumbre. Repite la adjudicación hasta agotar la cola. Estas decisiones quedan en `adjudication.sqlite`. Las etiquetas originales y su acuerdo se conservan. La falta de una segunda opinión sigue figurando como pendiente.

```
python3 review.py prepare --cases CASOS.json --events SUCESOS.json --freeze PROTOCOLO_FIJADO --freeze-sha256 HUELLA_REGISTRADA --mode human_collection --out LOTE
```

```
python3 review.py session A.sqlite
```

```
python3 review.py show A.sqlite
```

```
python3 review.py record A.sqlite --ticket CODIGO --label indeterminate --evidence "Falta precisar la fecha aplicable" --seconds 70
```

```
python3 review.py status A.sqlite
```

```
python3 review.py seal A.sqlite
```

```
python3 review.py report LOTE
```

```
python3 review.py adjudicate LOTE
```

```
python3 review.py adjudicate LOTE --case ID_CASO --label other --evidence "Explica la decisión y la fuente" --seconds 95
```

```
python3 review.py report LOTE
```

```
python3 -B -m unittest discover -s . -p test_review.py
```

Para un caso se usa `session` o la pareja `show` y `record`. La segunda persona utiliza su propia base en su dispositivo. Los segundos de los comandos son ejemplos. En los ensayos usa `--mode simulation` y conserva los lotes de simulación separados de las observaciones humanas.

Cómo publicar las mediciones

La matriz de tres por tres sitúa a A en filas y a B en columnas. El acuerdo bruto divide la diagonal por todos los casos con dos respuestas. La κ de Cohen sin ponderar es (acuerdo observado − esperado) / (1 − esperado). El acuerdo esperado se obtiene de los marginales de cada lector. Se conserva la categoría indeterminada, incluidas las coincidencias entre dos indeterminados. Sin parejas completas, acuerdo y κ son nulos. Cuando el acuerdo esperado es uno, κ queda sin definir aunque ambas personas hayan coincidido siempre. Publica la cobertura de parejas y los pendientes de cada persona junto a la matriz.

El informe JSON separa el esfuerzo por persona, por pareja y en adjudicación. Da los minutos-persona observados, la mediana y el percentil 90 de los segundos activos según el método del rango más próximo, y los recuentos de cronómetro y entrada manual. La duración por pareja suma los dos tiempos en casos con doble revisión. El esfuerzo total incluye una sola vez las revisiones completadas sin pareja y las adjudicaciones terminadas. Habría que medir aparte las esperas, los abandonos, la selección y la formación de personas, y el trabajo futuro. Conserva las resoluciones indeterminadas y las cantidades sin resolución final. Indica tamaño de la cohorte, idiomas y exposición a casos públicos.

Derechos y protección de datos en un despliegue

La organización debe determinar la base jurídica y la finalidad del tratamiento, minimizar lo que ve cada persona, limitar la conservación y controlar el acceso. Utiliza ejemplos sintéticos en público. Las notas o el historial de comandos pueden incluir datos personales. Introduce solo el contenido indispensable y revisa el riesgo de identificación antes de publicar agregados. Los códigos seudónimos pueden asociarse a casos mediante el fichero de correspondencias. El artículo 85 del Reglamento de IA permite reclamar ante la autoridad de vigilancia del mercado si hay motivos para considerar infringido el Reglamento. El artículo 86 reconoce, cuando se cumplen sus condiciones, el derecho a una explicación del responsable del despliegue sobre el papel de la IA y los elementos principales de la decisión. La propia resolución puede impugnarse por la vía administrativa o judicial aplicable. Los artículos 77 y 79 del RGPD regulan la reclamación en materia de datos y la tutela judicial. Una etiqueta de revisión no determina por sí sola el derecho a reparación ni la legalidad del despliegue.

Fuentes oficiales consultadas el 21 de noviembre de 2026. Reglamento de IA consolidado · Artículo 85 · Artículo 86 · RGPD, en particular artículos 5, 6, 32, 77 y 79.

Abrir la mesa de revisión

Necesitas tu propio paquete **A.sqlite o B.sqlite**, Python 3.10 o posterior y un navegador moderno. Cada persona trabaja en su ordenador. Desde el directorio que contiene la carpeta `v11`, abre una terminal. En el ejemplo, sustituye `A.sqlite` por la ruta al paquete recibido:

```
python3 v11/desk/desk.py A.sqlite --authored-holdout
```

En Windows se puede usar `py -3` en lugar de `python3`. Se abrirá el navegador. Si no aparece, copia en él el enlace privado que muestra la terminal. Deja la terminal abierta. La página lleva el código y los estilos incrustados. Se sirve en `127.0.0.1` con un puerto temporal, funciona sin Internet y no consulta las direcciones de las fuentes. El valor al final del enlace permite acceder al paquete en esa sesión. Guárdalo en privado. Utiliza una sola pestaña y evita abrir el mismo paquete a la vez con la herramienta de línea de comandos.

Una revisión de principio a fin

1. Elige **ES** o **EN** y pulsa **Empezar o reanudar**. El cronómetro se pone en marcha. Los tres paneles muestran la pregunta y la respuesta originales, las fichas de la petición si existe una captura, y el suceso con su fecha y la fuente retrospectiva. El idioma de la interfaz se puede cambiar sin modificar el texto evaluado.
2. Lee sin pistas. Puedes seleccionar un fragmento y pulsar **Subrayar selección**. Con teclado, abre **Seleccionar mediante teclado**, escribe un fragmento exacto y pulsa **Subrayar fragmento**. Si se repite, añade contexto para identificarlo. La marca continua corresponde a tus propias selecciones. Puedes quitar cada una. La pestaña no tiene acceso a etiquetas de referencia ni a predicciones del detector.
3. Elige la etiqueta y escribe por qué. El borrador se guarda al escribir. **Pausar** guarda el texto y detiene el cronómetro. Cambiar de pestaña solicita una pausa. El servidor recibe una señal cada cuatro segundos. Si desaparece sin pausarse, contabiliza como máximo quince segundos desde la última señal y exige reanudación. Al cerrar de forma abrupta se puede perder la última fracción de tiempo o texto aún pendiente de guardar. Pausa antes de salir. Abre el mismo paquete y pulsa **Empezar o reanudar** para continuar.
4. Pulsa **Guardar y continuar** cuando la decisión esté lista. La etiqueta, la justificación, las selecciones exactas y los segundos activos quedan en las mismas tablas que utiliza el kit de revisión. Las decisiones terminadas se conservan. La persona que adjudica sigue el protocolo de revisión.
5. Al completar todos los casos, pulsa **Cerrar paquete y obtener recibo**. Conserva el archivo SQLite y la huella indicada. Devuelve el paquete por el canal privado acordado y la huella por separado. Cierra la terminal con Ctrl+C. Para cerrar expresamente un paquete parcial, utiliza `review.py seal A.sqlite --allow-pending` según el protocolo.

Modo asistido. Solo se activa al pedirlo expresamente dentro de un caso. Subraya con una línea discontinua los años de cuatro cifras entre 1500 y 2199. Una coincidencia no establece que haya una afirmación ni que se incumpliera el corte. La activación se conserva aunque se pause o se cierre la página. Cada nueva respuesta empieza sin pistas. Si el protocolo exige doble ciego, conserva ese modo en todos los casos y declara cualquier exposición anterior. Para contar las dos condiciones por separado usa:

```
python3 v11/desk/desk.py A.sqlite --summary
```

El resumen distingue las decisiones ciegas, las expuestas a ayuda y las procedentes de otras interfaces. Los segundos asistidos incluyen solo los intervalos activos posteriores a la petición de ayuda. La nota sellada conserva el modo y los fragmentos de evidencia con sus posiciones en puntos de código Unicode. El informe general de acuerdo del kit combina las decisiones según el protocolo. Para interpretar un resultado ciego hay que apartar de ese análisis los casos asistidos y registrar la exposición previa. La pausa, las huellas y el cronómetro de la aplicación no prueban que dos personas trabajaran de manera independiente.

Qué recibió realmente el sistema

El indicador `--authored-holdout` corresponde al conjunto de evaluación redactado para la versión 1.1. No hubo una generación por un modelo ni envío de fichas durante aquella generación. Las referencias del tercer panel sirven para comprobar posteriormente el suceso. Para un paquete de origen desconocido, omite ese indicador y la interfaz dirá que falta una captura.

En un estudio con solicitudes registradas, coordinación puede adjuntar únicamente las capturas asociadas **a los códigos ciegos de ese lector**. Sustituye la opción anterior por `--contexts contextos_A.json`. Este JSON es un objeto cuyas claves son los códigos del paquete. Cada valor contiene `request_utf8`,

`request_sha256` y `output_sha256` de la captura. La petición codificada conserva `prompt`, `cutoff` y `sources`, con los campos de fuente `text`, `date` y `url`. La mesa comprueba ambas huellas y la correspondencia de la pregunta y el corte, y solo presenta las fichas contenidas en esa petición. Una lista vacía comprobada se muestra como «cero fichas». Si falta el código, figura «se desconoce». La correspondencia con el registro local no demuestra uso por un proveedor ni una relación causal. Revisa los textos y enlaces permitidos para evitar una revelación accidental. Usa únicamente las fuentes autorizadas y mínimas para esta decisión.

Qué exige la UE sobre los idiomas

El Reglamento de IA no establece una obligación universal de que todo sistema de IA responda en todas las lenguas oficiales de la UE. El artículo 13.2 exige instrucciones comprensibles, accesibles y claras para responsables del despliegue de sistemas de alto riesgo. El artículo 47.2 exige traducir la declaración UE de conformidad a una lengua que entiendan fácilmente las autoridades nacionales pertinentes. El artículo 50.5 exige que las comunicaciones de transparencia comprendidas en ese artículo sean claras, distinguibles y accesibles. Son obligaciones con destinatarios y objetos concretos. La lengua de un trámite público, los derechos lingüísticos nacionales y las necesidades de las personas usuarias deben examinarse para el servicio concreto. Esta herramienta permite revisar en español e inglés y conserva el idioma original de la evidencia.

Fuentes oficiales consultadas el 21 de noviembre de 2026. Reglamento de IA, texto consolidado · Artículo 13 · Artículo 47 · Artículo 50.

Verificación reproducible

```
python3 -B -m unittest discover -s v11/desk -p test_desk.py -v
```

Las pruebas usan casos simulados, tiempos de prueba y bases temporales. Comprueban pausa, persistencia, contextos, control de acceso local, modo asistido y el recibo de la base. Con Node 18 o posterior se ejecuta además el JavaScript entregado contra el servidor HTTP real con una representación del DOM de prueba. Ese recorrido incluye selección de evidencia, cambio de idioma, dos decisiones y cierre. La prueba no mide legibilidad visual ni valida el uso con lectores de pantalla. Cero participantes humanos.

¿Qué se puede fechar y qué se midió?

La edición 1.0 dejaba 58 referencias B sin fecha histórica suficiente. El 21 de noviembre de 2026 volvimos a consultar los 56 destinos consignados y a buscar las dos denominaciones sin enlace. Para **53 referencias** identificamos un objeto concreto que sí tiene una fecha defendible. En **27** casos esa fecha es anterior al 4 de diciembre de 2025 y en **26** es posterior. **Cinco** conservan abierta la cuestión de la fecha pública. Dos entradas M estudiadas aparte aportan una corrección adicional: hay una captura pública de ALETHEIA en agosto de 2025 y una fecha de registro en marzo de 2026 para el paquete Rust candidato. El recuento de 58 excluye estas dos entradas. Registro de las 58 referencias · 29 búsquedas y comprobaciones dirigidas

Cómo interpretar una fecha

Evidencia encontrada en las 58 fichas	Número	Qué fecha exactamente
Fecha que el editor atribuye al artículo	29	Se tomó de la página o de sus metadatos estructurados. Cuando aparecen publicación y modificación se conservan ambas en la ficha.

Evidencia encontrada en las 58 fichas	Número	Qué fecha exactamente
Copia archivada del destino	12	Acredita que se podía consultar la edición capturada en ese momento. Una captura posterior al corte deja abierta la posibilidad de una publicación anterior.
Publicación relacionada con edición o acto fechado	4	FINOS, V-Proof, Giza y Aleo tienen una publicación identificable. La fecha se aplica a ese borrador, entrevista, presentación o anuncio.
Únicamente fecha de revisión o actualización	4	B012, B034, B068 y B071. El sitio identifica una modificación, actualización de ficha o revisión del documento. El origen puede ser anterior.
Versión distribuida mediante un registro público	2	UQLM 0.1.0 y GRAIL v0.0.5. El registro vincula la fecha con la versión concreta.
Publicación administrativa o edición fechada	2	Una entrada en el índice de la SEC y una edición presupuestaria de la USPTO.
Sin fecha pública suficiente	5	Las limitaciones de identificación y recuperación se detallan más abajo.

La columna `object_date_vs_2025_12_04` compara la fecha del objeto identificado con la del corte. No es una clasificación jurídica ni cuenta primeras divulgaciones. Por ejemplo, B026 describe hoy una edición de julio de 2026, aunque la lista de mayo citaba otra con «2025» en el título. En B042, las fechas actuales de julio y agosto también son posteriores a la cita de mayo. En B044, el metadato de publicación es del 6 de mayo y la fecha de acceso registrada era el 5. Queda abierta la identidad de esas ediciones históricas. La edición 1.0 conserva el registro original de esas incertidumbres. B026, B042 y B044 en el registro

Cuatro pistas técnicas examinadas de nuevo

ALETHEIA (M37). La captura del repositorio en Internet Archive del **26 de agosto de 2025 a las 14:50:47 UTC** contiene el nombre del proyecto, el archivo de código y la descripción de un registro de decisiones encadenado por hash. GitHub conserva tres commits de ese día, incluido uno a las 13:56:22 UTC. Ahora hay evidencia externa de que esa descripción concreta era pública antes del corte. Es una referencia para trazabilidad de agentes. La captura no documenta una evaluación de disponibilidad histórica de hechos en respuestas generadas. Captura de agosto · Commit identificado

El módulo Rust (M38). La pista original decía `temporal_retrieval`. Identificamos y comprobamos el candidato concreto `graphrag-core`. El registro fecha la versión 0.1.0 el 12 de marzo de 2026 y la 0.1.1 el 18 de marzo. El archivo descargado de la segunda coincide con el SHA-256 publicado, `dc86967bf8af1084262f5c481521585619a37ec1cb023479c0e02cbb0c0bef4d`. El paquete contiene 233 archivos, incluido `src/graph/temporal.rs`, y documenta intervalos temporales opcionales en entidades y relaciones. La búsqueda literal en sus archivos Rust, Markdown y TOML no encontró `temporal_retrieval`. Esta evidencia permite comparar tipos e intervalos en una versión posterior. La identidad de la fuente histórica y la existencia del supuesto filtro obligatorio antes del corte siguen abiertas. Registro de versiones · Versión y checksum · Documentación del paquete

GRAIL (B093). La denominación coincide con *Guaranteed Rollout Authenticity via Inference Ledger*. GitHub registra v0.0.5 el **15 de septiembre de 2025 a las 21:40:22 UTC**. La captura independiente del **7 de octubre** conserva la denominación desarrollada y explica que el proyecto vincula una trayectoria generada durante entrenamiento por refuerzo con un modelo y una entrada. La referencia de mayo carecía de enlace, por lo que la atribución histórica sigue siendo inferida de la coincidencia del nombre y de la descripción. La captura demuestra que ese uso técnico del término era público antes del corte. El predicado relevante es la

autenticidad de una trayectoria, no la fecha en la que un hecho llegó al dominio público. Lanzamiento · Captura del repositorio

MECM (B094). Se buscó el acrónimo junto a *temporal causal gating*, *autoencoder* y términos de razonamiento causal. Aparecieron modelos de razonamiento emocional y de redes cerebrales. Ningún candidato permitió identificar al autor, el título o una descripción que correspondiera inequívocamente a la pista. No se asigna fecha ni función metodológica. Búsquedas S04, G6 y U6

Otras correcciones que cambian la lectura

- **Incertidumbre y gobernanza anteriores al corte.** PyPI fecha los artefactos de UQLM 0.1.0 el 6 de mayo de 2025. Una captura de Edimburgo del 15 de noviembre de 2025 conserva un índice público de investigación sobre alucinaciones. El primer borrador de FINOS se anunció el 1 de octubre de 2024 y la versión 1 en junio de 2025. Las tres evidencias sirven para comparar una distribución concreta, un índice y una guía de gobernanza. Los cambios posteriores se fechan por separado. UQLM · Índice archivado · FINOS, borrador · FINOS v1
- **Productos y afirmaciones públicas.** Una presentación de Giza de 2023 y el anuncio de la red principal de Aleo en 2024 documentan descripciones previas. Una entrevista del 13 de mayo de 2026 describe V-Proof y su material comercial actual explica la verificación de evidencias. Es posible fechar la información publicada por la prensa. Quedan por acreditar la primera versión técnica y su funcionamiento. Giza · Aleo · V-Proof
- **Carta fechada y publicación pública.** En B066, la carta particular lleva fecha de 5 de octubre de 2025. El índice de aportaciones de la SEC data la entrada el 14 de diciembre. Por ese canal consta una aparición posterior al corte. La circulación pública anterior y las fechas de las referencias incluidas en la carta requieren otras pruebas. La carta expresa la posición de quien la presenta. Carta · Índice de la SEC

En B013 se identifican las diapositivas de Alexey Zaytsev y una fecha de creación/modificación del PDF del 21 de abril de 2025, pero falta una prueba de publicación pública en esa fecha. B033 sigue siendo una página corporativa sin edición histórica fechada. B051 y B055 son consultas dinámicas: el resultado actual y un fallo de recuperación no permiten reconstruir la lista vista en mayo. B094 es la denominación MECM aún sin resolver. Incluso entre las 53 fichas con una fecha subsisten límites de primera publicación, versión y correspondencia que se especifican por fila. Registro completo

S(t), A(t) y Δ(t) como hipótesis de medición

La hipótesis original propone $S(t) = \alpha C + \beta E + \gamma R$. C representaría acuerdo entre ejecuciones o modelos bajo controles definidos. E expresaría estabilidad ante perturbaciones de la instrucción o del muestreo. R mediría el reconocimiento de un concepto en un espacio de representaciones previamente fijado. Estimar esos términos exige especificar modelos, representaciones, distribución de perturbaciones, repeticiones, unidad de análisis, normalización, pesos y variabilidad. La densidad semántica de Qiu y Miikkulainen y la entropía semántica de Farquhar y colaboradores ofrecen comparadores para la incertidumbre de respuestas. El trabajo posterior sobre información por token estudia otra magnitud. Ninguno proporciona los valores de C, E y R para esta colección. Qiu y Miikkulainen, v1 · Farquhar y colaboradores · Densidad de información por token

A(t) designaría el respaldo observable de una proposición en un corpus identificable disponible hasta t. Una estimación tendría que definir qué se cuenta como respaldo, qué fecha determina la inclusión de un documento, cómo se tratan correcciones y fuentes duplicadas y cuándo la cobertura se considera insuficiente. La fecha de ocurrencia de un suceso, la primera noticia y la fecha en que un índice incorporó ese documento pueden diferir. Sin una escala comparable para S y A, el signo de $\Delta(t) = S(t) - A(t)$ depende de

decisiones de normalización. Una alerta débil requeriría términos definidos y calibrados y una evaluación separada de su incertidumbre. Estudio comparable de inferencia ex ante

La evaluación publicada cuenta etiquetas, decisiones y abstenciones de clasificadores sobre respuestas redactadas y sucesos registrados. El comparador de años explícitos recibe un indicador externo sobre el contexto. El clasificador de proposiciones v1.1 aplica indicios textuales y fechas suministradas por el auditor. Su umbral, seleccionado con ejemplos de desarrollo, atañe a la decisión de emitir una etiqueta o solicitar revisión. **No se estimaron S(t), A(t), sus pesos ni curvas de Δ(t)**. Por ello, los resultados no permiten informar una detección calibrada de activaciones no causales ni atribuir una respuesta al entrenamiento, a la recuperación o al azar. Detector de años en la revisión pública · Especificación del clasificador · Umbral publicado

Qué usar y cómo citar

Para una auditoría o reclamación, empezaríamos por la captura local, la versión de cada fuente y la posibilidad de reabrir un caso. Reservamos la alerta lingüística para apoyar esa revisión. Con el conjunto publicado, un sistema que cerrase expedientes por su cuenta cometería errores incluso al abstenerse con frecuencia. Los siguientes datos decisivos son tiempos y desacuerdos de personas independientes sobre textos auténticos. El ensayo piloto, el reto de seis sucesos, las traducciones, el lector con IA y los canarios tienen unidades distintas. No cabe agregarlos como si fueran una sola tasa de acierto.

El autor desarrolla la tecnología evaluada. El software se distribuye por separado bajo Business Source License 1.1. La edición completa incorpora a continuación el trabajo de base y su catálogo de antecedentes. Investigación, desarrollo, traducción y edición recibieron asistencia de IA.

Rubén Abella (2026). *Temporal auditing of AI outputs: request-level evidence, prior art and the limits of automated detection*. SPE Audit, v1.1. DOI 10.5281/zenodo.24424024 · GitHub

Reutilización y atribución

© 2026 Rubén Abella. El texto y las figuras originales de esta publicación se ofrecen con la licencia Creative Commons Atribución 4.0 Internacional. Al reutilizarlos, debe reconocerse a Rubén Abella, enlazarse la licencia e indicarse los cambios. La autorización comprende únicamente el texto y las figuras originales del autor. Los materiales de terceros conservan sus propias condiciones. El nombre SPE Audit, su logotipo y las demás marcas quedan excluidos de esta autorización. Al software le corresponden las condiciones de la Business Source License 1.1 del repositorio. CC BY 4.0 · Licencia del software

SPE Audit

ESTUDIO ABIERTO

Auditoría temporal de las respuestas de IA

Evidencia por petición, antecedentes técnicos y límites de la detección automática

Rubén Abella · SPE Audit

21 de noviembre de 2026

EDICIÓN DE INVESTIGACIÓN 1.0 · ES

Índice

Resumen . . . 3

Una pregunta que se pueda responder en público . . . 3

Qué puede demostrar un registro de auditoría . . . 5

Evidencia y resultados . . . 7

Cómo medir una afirmación sobre el pasado . . . 11

Qué puede preguntar la ciudadanía al amparo del Reglamento de IA de la UE . . . 13

Cómo se sitúa este estudio entre las revisiones publicadas . . . 15

Revisión de noviembre. Qué se sabía sobre auditoría temporal de la IA . . . 16

Conclusiones . . . 22

Los 54 registros de la comparación de mayo . . . 23

Cobertura de la revisión y procedencia de las referencias . . . 32

Catálogo de las 116 referencias reconciliadas y complementarias . . . 37

Resumen

Una respuesta puede describir correctamente algo que ocurrió meses después de la fecha desde la que se pidió contestar. Para quien recibe esa respuesta surgen dos preguntas diferentes. ¿Presenta como consumado un hecho todavía futuro? ¿Qué evidencia conserva la aplicación acerca de cómo produjo el texto? Estudiamos ambas preguntas mediante la revisión de antecedentes, el examen de la implementación pública de SPE Audit y un experimento reproducible sin credenciales. Nuestro criterio positivo es una afirmación de un desenlace posterior al corte expresada como hecho consumado. Una alerta clasifica el texto. Atribuir ese texto a una fuente concreta requiere evidencia adicional.

Comparamos el detector anual público con una regla que usa fechas de sucesos, lenguaje y abstención explícita. En 96 casos construidos, el primero detectó 4 de 32 positivos y produjo 12 falsas alarmas. La segunda regla detectó 24 y remitió 12 casos a revisión. En 16 frases redactadas por separado detectaron, respectivamente, 1 y 0 de los 8 positivos. La segunda regla remitió 14 de los 16 a revisión. Doce intervenciones sobre el contenido recuperado muestran una dependencia comprobable en un emisor local determinista. Una revisión separada conserva los 25 intentos de un piloto anterior, incluidos tres errores de servicio. Publicamos cada predicción, las reglas, las fichas de evidencia y un verificador de las 124 capturas del estudio controlado.

La revisión recupera el inventario histórico, la bibliografía de los dos informes de origen y los comparadores añadidos en noviembre. Los trabajos más próximos ya cubren recuperación histórica, evaluación de información futura y procedencia criptográfica. La contribución de este estudio es hacer examinable su combinación a escala de una petición y medir lo que se pierde cuando cambia la redacción. La contribución al debate europeo es práctica. Una persona puede comprobar qué texto se registró, discutir la clasificación y preguntar por la evidencia que falta.

Una pregunta que se pueda responder en público

Imaginemos que una persona pide a un asistente que reconstruya la situación de una misión espacial a 1 de enero de 2022. En aquel momento ya se había anunciado la intención de que DART chocara con Dimorphos. La NASA comunicó el impacto en septiembre. «Está previsto un impacto en 2022» y «la nave impactó en el asteroide» comparten tema, vocabulario y aparentemente una misma fecha. Solo la segunda presenta como ocurrido el desenlace. Si una pantalla marca en rojo la cifra 2022, ¿puede quien recibe la alerta saber por qué? Si no hay alerta, ¿puede saber si el detector entendió el verbo? Esta diferencia afecta a cualquier reconstrucción histórica, desde una noticia hasta la explicación que recibe quien reclama una decisión. Anuncio del lanzamiento · Imágenes y confirmación del impacto

La desconfianza en sistemas de IA abarca cuestiones de poder, derechos, supervisión y calidad de las decisiones. Aquí formulamos una pregunta más pequeña a la que sí podemos dar una respuesta verificable. ¿Puede otra persona inspeccionar la petición que una aplicación preparó para una fecha, leer las fuentes que entregó y poner a prueba la interpretación de la respuesta? El expediente mínimo sirve a tres lectores. La persona afectada necesita una explicación comprensible y un mecanismo de rectificación. El operador necesita recuperar la ejecución concreta. El investigador necesita repetir la prueba sobre el mismo material y conservar los desacuerdos.

Qué cuenta como resultado

1. **Evidencia de contenido.** Una ficha del estudio identifica fuente, fecha declarada y texto. La petición conserva las fichas exactas entregadas a una función local. Al cambiar un byte cambia la huella que espera quien verifica.
2. **Interpretación temporal.** La etiqueta se aplica a la afirmación realizada por el hablante. Una fecha futura puede aparecer legítimamente en un plan, una pregunta o una negación.
3. **Capacidad operativa.** Una regla que se abstiene en los casos difíciles impone trabajo al revisor. Publicamos el número de positivos detectados entre todos los positivos y la cobertura sobre todas las filas.
4. **Atribución.** El archivo local permite identificar lo que el cliente observó. Para atribuir al proveedor una entrada, una fecha o una vía causal de acceso a información, el revisor necesita observaciones en ese punto.

Un contraejemplo establece el límite de la última inferencia. Dos servicios reciben la misma petición y responden con la misma frase. El primero reproduce una previsión que por azar resulta cierta. El segundo recupera fuera de la aplicación una noticia posterior. Los bytes y las huellas que ve el cliente pueden ser idénticos. Cualquier procedimiento que reciba exclusivamente esa evidencia debe devolver el mismo resultado en ambos mundos. Sin una observación que los distinga, su conclusión acerca del origen será incorrecta en al menos uno. Es una propiedad de la información disponible, incluso con una función de resumen criptográfico perfecta.

Cómo se construyó la revisión

Partimos de las posiciones numeradas del estudio histórico y de las dos listas de referencias conservadas. Para cada registro buscamos una identidad documental, una fecha y la operación técnica que realmente describe. Damos preferencia a documentos publicados, artículos en su versión identificada, especificaciones y expedientes de patente. Una versión posterior puede aclarar el diseño sin probar que el detalle estaba en una versión previa. La fecha de corte histórico es el **4 de diciembre de 2025**. La revisión se cerró el **21 de noviembre de 2026**. Conservamos duplicados entre índices cuando explican por qué se estudió una misma fuente desde dos preguntas distintas.

El lector encontrará tres vistas del material. Las filas M conservan el inventario de mayo en su orden original. Las fichas N profundizan en los trabajos más próximos y en cambios documentales posteriores. El catálogo B permite localizar las referencias que aparecieron en los informes y las fuentes suplementarias. Las listas se solapan. Una ficha puede enlazar solicitud, concesión y publicación internacional de una misma familia. Las pistas no resueltas siguen visibles. Este método recupera un corpus identificable, con búsquedas complementarias dirigidas por rasgos técnicos. No producimos un censo mundial ni una opinión jurídica sobre concesión o libertad de operación.

Cómo leer los resultados sin conocimientos técnicos

La palabra **positivo** significa que la frase presenta como ocurrido un hecho futuro respecto del corte. **Revisión** significa que una persona aún tiene que decidir. Para una primera lectura basta con abrir la demo, escoger DART, mirar la fecha, leer la ficha admisible y comparar qué decide cada regla. A continuación se puede verificar la huella y leer el razonamiento del caso. Quien quiera comprobar el informe completo puede ejecutar el mismo ensayo con Python y abrir un CSV con todas las discrepancias.

Qué puede demostrar un registro de auditoría

Una persona reclama por una respuesta que, según el servicio, se elaboró con la información disponible en una fecha concreta. La reclamación plantea varias preguntas. ¿Cuál es el texto impugnado? ¿Qué documentos entregó la aplicación? ¿Cuándo apareció la primera evidencia del hecho discutido? ¿Procesó el proveedor exactamente esa entrada? Resolver las dos primeras permite investigar las siguientes. Cada conclusión necesita una observación propia.

El motor público y el punto de captura

La versión de febrero contiene funciones para filtrar registros por fecha, ordenar sus descriptores, calcular la huella de una lista serializada, crear una cápsula y encadenar cápsulas en SQLite. La ordenación considera identificador documental, marca temporal e identificador de fuente. En esa versión, el campo llamado `context_merkle_root` puede contener una sola huella SHA-256 de la serialización completa. Su verificación requiere esa serialización. La implementación carece de una ruta de árbol con la que demostrar de forma independiente la pertenencia de una sola hoja. Núcleo público en la revisión estudiada

El documento técnico público ilustra una integración RAG completa con `context_items`. En la revisión fijada, la función pública `generate_proof` recibe texto, un archivo o una huella ya calculada. Asigna al contexto una huella fija y deja vacío el hash del *prompt*. El archivo ZIP generado contiene la cápsula, el registro SQLite y el manifiesto de entrada. El verificador comprueba el encadenamiento y la huella de la última cápsula. Si recibe además un archivo original, compara su huella. El indicador global de validez contempla la cadena, la vinculación de la cápsula y las discrepancias de archivo. El resultado de la firma se comunica por separado. Un expediente sin firma puede superar correctamente las comprobaciones de integridad descritas. Documento técnico · API pública

El estudio incorpora un adaptador identificado de forma independiente. Cada captura corresponde a una llamada a una función de envío. La entrada es un documento JSON en UTF-8 que contiene la pregunta, la fecha de corte inclusiva, la versión de la política y el texto y los metadatos de las fichas admisibles suministradas. El adaptador descarta fichas con fecha declarada posterior al corte, ordena las demás por identificador único y entrega al emisor exactamente esos bytes. Conserva el texto devuelto por la función. Calcula una huella de la petición y otra de la respuesta. Una tercera huella vincula ambas con el esquema y la política. El verificador vuelve a interpretar la petición, comprueba las fechas y la ordenación canónica, recalcula la huella del registro y admite una comparación con una huella esperada conservada de forma independiente. Otra huella, a escala de colección, fija los bytes exactos del archivo JSONL y su orden.

En el ensayo controlado, la función emisora reproduce la respuesta del caso. Si una integración posterior traduce la entrada a una petición HTTP, incorpora instrucciones, recupera más documentos o consulta otro modelo, la observación de este estudio termina en la entrada del emisor local. Un identificador de petición correlacionado con registros del proveedor aportaría evidencia adicional de aquella ejecución. La conservación independiente del fondo documental consultable, las reglas de selección, las versiones y sus fechas de adquisición permitiría investigar omisiones. Otra vía de consulta u otra versión de la aplicación constituye una ejecución distinta.

Conclusiones al alcance del revisor

Pregunta	Observación disponible	Evidencia adicional para ampliar la conclusión
¿Coincide la respuesta con el registro?	Comparar sus bytes y su huella. Recalcular el sello del registro y cotejarlo con el esperado.	Para atribución y fecha, una copia independiente, un firmante de confianza o un testigo temporal pertinente.
¿Cumplía cada fuente el corte?	Interpretar las fechas declaradas, aplicar el corte e inspeccionar el texto seleccionado.	Instantáneas históricas o archivos del editor que acrediten la disponibilidad de esa versión. Una página puede cambiar después de su publicación inicial.
¿Se utilizó todo el contexto previsto?	Examinar todas las fichas que cruzaron el punto de envío local observado.	Inventario y registro de recuperación para examinar la exhaustividad. Registro de recepción para verificar el procesamiento, incluidas instrucciones internas o llamadas a herramientas cuando estén disponibles.
¿La respuesta afirma que ocurrió un hecho posterior?	Identificar suceso, afirmación, carácter hipotético, negación y evidencia fechada de contraste.	Un revisor o un conjunto adjudicado puede resolver ambigüedades y valorar qué era cognoscible en la fecha.
¿De dónde obtuvo el modelo la información?	Una incoherencia puede motivar que se abra una investigación.	Evidencia que distinga pistas de la pregunta, fuentes recuperadas, asociaciones aprendidas, predicción razonable, coincidencia y otros canales.

La política temporal debe indicar a qué fecha se refiere. Un contrato firmado un lunes, registrado un jueves e incorporado al buscador la semana siguiente tiene tres fechas útiles. Para reconstruir el conocimiento público interesa la primera disponibilidad documentada de esa versión. Para investigar la información en poder de una organización puede importar el momento de ingesta. La elección cambia qué elementos resultan admisibles. El adaptador acepta el día declarado de publicación con precisión de calendario. No distingue una publicación a las 08:00 de una decisión a las 16:00 de ese mismo día. Las fichas del experimento son resúmenes preparados para este estudio que remiten a fuentes fechadas. Los bytes históricos del editor requerirían un archivo distinto.

Las medidas semánticas propuestas

El diseño original planteó $S(t) = \alpha C + \beta E + \gamma R$. C describiría el acuerdo entre ejecuciones controladas o entre modelos. E expresaría la estabilidad ante cambios en la pregunta o el muestreo. R mediría la densidad de reconocimiento de un concepto en un espacio de representaciones especificado. Para reproducir una estimación hay que publicar versiones de modelo y representaciones, variantes de las preguntas, tamaños de muestra, funciones de distancia, escalas y pesos. Si un modelo repite de forma sistemática una afirmación sin respaldo, puede puntuar alto en acuerdo y estabilidad. Esas mediciones describen un comportamiento.

La serie externa $A(t)$ pretendía representar el apoyo normalizado de fuentes fechadas y disponibles hasta t. Una aplicación necesita definir el universo de fuentes, la unidad contabilizada, el tratamiento de duplicados y revisiones, la cobertura y la versión y escala de normalización. El acuerdo entre respuestas de una cohorte anterior de modelos mide otra magnitud. Una búsqueda vacía en un archivo de escasa cobertura aporta evidencia limitada sobre lo que se sabía en el mundo. La diferencia $\Delta(t) = S(t) - A(t)$ adquiere sentido en la medida en que las escalas y ventanas de observación permitan compararlas.

La señal *fuerte* propuesta requiere una afirmación sobre un hecho posterior de fecha verificada y un contexto atestiguado anterior al corte. La heurística pública recibe del exterior un indicador sobre el contexto y busca años explícitos. Deja abiertas las cuestiones de negación, cita textual y pronóstico. La señal *débil* propuesta usa un $S(t)$ alto y un $A(t)$ bajo, con umbrales definidos. Interpretarla exigiría calibración

sobre ejemplos separados de la evaluación y un tratamiento para la evidencia insuficiente. El estudio diferencia el aviso de clasificación, la comprobación de integridad local y la explicación causal. No estima curvas de S/A ni presenta resultados de una señal débil calibrada. Detector público · ExAnte · Pitfalls in Evaluating Language Model Forecasters

Dos ejemplos para una revisión

Una misión prevista. Antes de aterrizar, el anuncio de una misión ya puede fijar para el año siguiente una fecha de aterrizaje. «El aterrizaje está previsto para el próximo año» y «la nave aterrizó con éxito» sostienen cosas distintas aunque mencionen la misma fecha. Conviene reunir la fuente disponible al corte, el comunicado posterior con el resultado y la pregunta exacta. Si la pregunta ya revelaba el desenlace, documentarlo como posible origen de esa información. El hallazgo aprovechable es distinguir con evidencia la previsión, el resultado y la repetición de una premisa.

Una reclamación por una decisión sobre una prestación (ejemplo hipotético). Una residente impugna una notificación que resume una decisión del 10 de marzo. Una frase menciona una modificación publicada por primera vez en abril. El revisor puede cotejar la notificación con la salida conservada y examinar el texto fechado de cada ficha de la captura local. Si la respuesta atribuye la decisión de marzo a la modificación de abril, se puede señalar la contradicción exacta y solicitar la versión documental, el expediente de decisión y el historial de aprobación humana. Si ya existía en marzo un borrador con esa misma previsión, la valoración cambia. La auditoría de la explicación por sí sola deja abierta la cuestión de qué regla aplicó quien tomó la decisión. El supuesto ilustra cómo ordenar evidencias, sin pronunciarse sobre ninguna administración.

Estos ejemplos pueden reproducirse públicamente con documentos públicos o construidos. Los casos operativos pueden contener preguntas privadas, decisiones e identificadores. El formato del estudio conserva intencionadamente los bytes completos para el revisor. Quien lo utilice en un servicio deberá definir acceso, conservación y canal de entrega. Publicar una huella no anonimiza un valor corto y predecible. Un revisor autorizado puede recibir el original en un entorno controlado, y el ejemplo público puede utilizar registros redactados expresamente e identificados con su procedencia.

Evidencia y resultados

Cuatro pruebas responden a cuatro preguntas diferentes. Los 96 casos construidos comparan dos reglas sobre frases en inglés cuyo contenido conocemos. Un segundo conjunto de 16 casos permite observar qué ocurre al cambiar la redacción. Doce intervenciones muestran cómo responde un lector automático local cuando cambian los documentos que recibe. Por último, un registro anterior conserva 25 peticiones remotas con una fecha de corte histórica. Las dos últimas pruebas tienen unidades de análisis distintas y no aumentan el tamaño de la muestra con la que se evalúan los detectores.

Qué se midió

La etiqueta positiva significa que una respuesta da por consumado el suceso objetivo aunque ese suceso ocurrió después de la fecha de corte, que se toma como inclusiva. La etiqueta describe lo que afirma la frase. Un acierto por casualidad, una pregunta que anticipa la respuesta, una invención que coincide con el desenlace, un documento recuperado y el propio entrenamiento pueden producir la misma frase. El texto aislado no permite distinguir esas causas. La etiqueta negativa incluye un pronóstico, una pregunta, una negación o el recuerdo de un hecho que ya había ocurrido en la fecha pedida. Se refiere a la proposición evaluada. Otras afirmaciones de la misma respuesta podrían contener errores.

Elegimos cuatro sucesos con una descripción anterior del plan y otra fuente que documenta su realización. Cada ficha del estudio es un resumen editorial breve con un enlace. El lector puede inspeccionar los bytes que recibió el código local. Una ficha no acredita por sí sola qué mostraba una página web dinámica en una fecha pasada.

Misión	Fecha de la ficha inicial	Suceso	Fuente elegida sobre el desenlace
Perseverance	30 julio 2020	Aterrizó en Marte el 18 de febrero de 2021	NASA, 18 febrero 2021
DART	24 noviembre 2021	Impactó contra Dimorphos el 26 de septiembre de 2022	NASA, 27 septiembre 2022
OSIRIS-REx	10 mayo 2021	Entregó en la Tierra la muestra de Bennu el 24 de septiembre de 2023	NASA, 24 septiembre 2023
Rosetta	3 junio 2011	Despertó de la hibernación el 20 de enero de 2014	ESA, 20 enero 2014

El detector público extrae años de cuatro cifras entre 1900 y 2099. Si se le indica que el contexto carecía de información posterior al corte, emite una alerta fuerte cuando encuentra un año mayor que el año límite. Usamos el detector y el extractor originales de una revisión identificada del repositorio público. El segundo método recibe la identidad y la fecha del suceso objetivo. Compara expresiones fijas para reconocer previsiones, negaciones, preguntas y determinados enunciados de hechos consumados. Cuando no reconoce la redacción, devuelve «revisión». Conocer la fecha exacta del suceso es una ventaja informativa. Las diferencias observadas incluyen tanto esa ventaja como el distinto tratamiento del lenguaje. Detector público de SPE

Las 96 respuestas construidas

Cada misión aporta ocho formas de respuesta que se reproducen ante tres fechas de corte: fin del año anterior, 1 de enero del año del suceso y fin del año del suceso. Las ocho formas son una afirmación con fecha, una afirmación sin fecha, una paráfrasis, un pronóstico, una negación, una pregunta repetida, una abstención y un enunciado mixto. La misma frase vuelve a evaluarse al mover el corte. Antes del suceso, cuatro formas por misión lo dan por realizado. Después, las mismas palabras pueden describir válidamente un hecho pasado. Así se obtienen 32 positivos y 64 casos de otro tipo. Son cuatro sucesos con observaciones relacionadas entre sí; contar 96 filas no equivale a haber observado el comportamiento independiente de 96 modelos o usuarios.

Cada reproducción guarda la petición exacta sometida al filtro local y la respuesta escrita de antemano. La función que devuelve esa frase prescinde del contenido de los documentos. Ninguno de estos 96 ensayos consulta un modelo remoto. El programa calcula las predicciones de ambos métodos antes de leer las etiquetas de evaluación. Así se evita entregar al detector la respuesta correcta como argumento. Sigue existiendo una relación entre el comparador escrito a mano, las misiones elegidas y los ejemplos de desarrollo.

Método y conjunto	Alarma correcta	Falsa alarma	Otro correcto	Positivo clasificado como otro	Revisión, positivo / otro
Regla anual, 96	4	12	52	28	0 / 0
Comparador con fechas, 96	24	0	60	0	8 / 4

Método y conjunto	Alarma correcta	Falsa alarma	Otro correcto	Positivo clasificado como otro	Revisión, positivo / otro
Regla anual, 16	1	2	6	7	0 / 0
Comparador con fechas, 16	0	0	2	0	8 / 6

La regla anual decidió los 96 casos. Detectó 4/32 positivos (sensibilidad del 12,5 %) y acertó en 4/16 alertas emitidas (precisión del 25 %). Emitió 12 falsas alarmas entre los 64 negativos. Las doce correspondían a un año futuro dentro de una previsión, una negación o una pregunta repetida en la condición con corte en el año anterior. Entre los 16 positivos anteriores al suceso, pero ya dentro del mismo año, detectó cero. Al mover el corte al mes de enero cambió el resultado de la comparación numérica entre años, aunque las cuatro misiones seguían pendientes de su desenlace.

El comparador con fechas decidió 84/96 casos (cobertura del 87,5 %). Detectó 24/32 positivos (75 %) y no generó falsas alarmas entre los 64 negativos. Sus 24 alertas coincidieron con la referencia. Los otros ocho positivos fueron a revisión. Si se miran solo los casos decididos se obtienen 84/84 aciertos. Ese resultado depende de que el algoritmo reconozca la formulación. Las paráfrasis que dejó pendientes ya permiten identificar parte del trabajo que recaería sobre el auditor. El segundo conjunto mide cuánto cambia esa carga con otra redacción.

El desafío de 16 casos con redacción nueva

En una fase de redacción separada se prepararon cuatro frases nuevas por misión, con ocho positivos en total. Las preguntas no revelan una fecha numérica. Quien redactó las frases conocía la pregunta de investigación y la limitación general de una regla basada en años. Las respuestas y sus etiquetas quedaron fijadas antes de consultar durante esa fase el comparador o el conjunto principal. El programa comprueba las huellas del comparador, del desafío y de sus etiquetas. Una alteración posterior respecto de esos ficheros queda así identificada. El procedimiento tampoco incorpora una revisión científica externa, doble anotación ciega ni sucesos nuevos.

La regla anual decidió los 16 casos, reconoció 1/8 positivos y activó dos falsas alarmas entre los ocho controles. Una corresponde a una frase que prevé que Rosetta despierte en 2014, condicionándolo a que funcione su temporizador. El comparador con fechas solo decidió 2/16 (cobertura del 12,5 %), ambos controles. Derivó 14/16 a revisión (87,5 %), incluidos los ocho positivos. Su sensibilidad automática fue 0/8. Al no emitir alarmas, su precisión entre alarmas no está definida. No se midió qué habría resuelto un revisor humano con esos 14 expedientes.

El cambio de redacción puede parecer pequeño al leerlo. El comparador reconoce la expresión empleada en los ejemplos principales para indicar que un róver aterrizó en Marte. En el caso h16, la respuesta dice que Perseverance se construyó para aterrizar y que «lo hizo» en febrero. Una persona puede identificar a qué se refiere «lo hizo» y contrastarlo con el corte del 10 de febrero de 2021. El comparador se abstiene y la regla anual omite la afirmación, ya que 2021 es también el año del corte. Esta pérdida de cobertura, de 84/96 decisiones en el conjunto construido a 2/16 en el desafío, muestra cuánto trabajo manual puede ocultar una evaluación con expresiones conocidas. Mezclar los dos grupos no daría una estimación de prevalencia en producción ni una clasificación fiable de modelos.

Doce intervenciones sobre los documentos entregados

Para cada misión, una función local recibió fichas seleccionadas por fecha. Su regla consiste en devolver el texto de la ficha admitida más reciente, desempatando por identificador. En la primera condición se ofrecen al filtro las dos fichas, pero solo la inicial resulta admisible. En la segunda, ambas cumplen el corte. En la tercera se mantiene ese corte posterior y se retira la ficha de desenlace del conjunto candidato. Se puede inspeccionar el cambio en la entrada de un componente que está bajo control del experimento.

Las cuatro misiones produjeron la misma secuencia: plan, desenlace, plan. Para DART, la primera respuesta habla de la intención de provocar un impacto. Al admitir la segunda ficha se informa de que el impacto ocurrió el día anterior. Al retirarla, con el mismo corte del 27 de septiembre, vuelve el plan. Las doce verificaciones acreditan cómo se conectan el filtro local, los bytes de la petición y esa función determinista. No miden el rendimiento de un modelo probabilístico, ni bajo ataque, ni acreditan qué información causó una respuesta en un modelo remoto.

El archivo de resultados contiene 124 registros verificados localmente: 96 reproducciones principales, 16 del desafío y 12 intervenciones. El CSV y el fichero de intervenciones indican para cada caso su línea y su huella. La versión de febrero de SPE incluye un registro SQLite que encadena huellas de cápsulas. La serie portátil de este experimento usa otro formato. Guarda petición local canónica, texto devuelto, SHA-256 por registro y una huella de los bytes del fichero JSONL completo en su orden original. Con una copia de esa última huella obtenida por una vía independiente, un revisor puede detectar cambios en el contenido o en el orden. Si alguien edita el archivo y vuelve a calcular la huella, la autoconsistencia no establece autoría ni fecha. Para esos registros locales no se obtuvo un sello de tiempo externo ni firma del proveedor.

El registro de aplicación de mayo

El registro histórico lleva fecha de 10 de mayo de 2026. Recoge cinco configuraciones y cinco preguntas en inglés. La instrucción exigía responder con corte de 1 de junio de 2023, sin documentos recuperados. Se conservaron 22 textos y tres errores HTTP 429. Las respuestas muestran dificultades que un conjunto construido podría pasar por alto. Una pregunta daba por supuesta una sentencia; una negación repetía parte de su vocabulario; otra respuesta reconocía actuaciones posteriores; y una descripción de un incidente aéreo situaba un suceso distinto en un año anterior. No hay acuses del servicio que autentiquen independientemente las revisiones de los modelos o el contexto en el servidor. Estos 25 intentos históricos son independientes del recuento de 124 capturas locales.

La regla antigua por palabras clave señaló uno de los 22 textos. Exigía dos términos específicos del suceso **y** que el evaluador ya hubiera etiquetado el suceso como posterior. Al agregar por suceso se asignaba a todos los controles anteriores una categoría que computaba como negativo correcto. La única respuesta señalada decía que el acusado aún no había recibido sentencia en la fecha de corte. La pregunta ya interrogaba por la pena definitiva de prisión. Otra respuesta afirmó que el juicio y la sentencia ocurrieron después, sin indicar la pena. Ninguna de las dos reglas la señaló. Al volver a aplicar el detector público por años a los 22 textos completos se obtuvieron cero alertas. Esos recuentos describen el registro y no bastan para estimar precisión, sensibilidad ni qué información pudo conocer legítimamente una configuración concreta.

El extracto público conserva términos detectados, años explícitos, citas breves de cuatro casos, paráfrasis editoriales y la huella de cada texto completo. Permite rehacer los recuentos desde los rasgos publicados. Volver a extraer todos los rasgos desde los textos íntegros requeriría el registro original. El ejemplo bancario da una pauta para interpretar los controles: el acuerdo entre UBS y Credit Suisse en marzo de 2023 y el cierre jurídico en junio son hechos distintos. Al etiquetar el tema entero como anterior al corte se pierde esa diferencia. Anuncio de FINMA · Comunicado de cierre de UBS

El resultado empírico conjunto tiene un alcance concreto. Un tercero puede reconstruir qué registró un emisor local controlado. Algunas alertas pueden refutarse leyendo la respuesta junto con su pregunta, y el comparador que publicamos necesita mucha revisión humana al cambiar la redacción. Identificar de dónde procede el conocimiento de un modelo alojado y medir el rendimiento en producción siguen siendo preguntas empíricas abiertas.

Cómo medir una afirmación sobre el pasado

La unidad de evaluación es una proposición sobre un suceso concreto. Ante «¿qué ha conseguido la misión?», decir que el róver llegó sano y salvo a la superficie afirma que el aterrizaje ocurrió, aunque la frase no dé el año. Un comunicado de lanzamiento que anticipe el día exacto de llegada expresa otra cosa. Para alguien que quiere impugnar una respuesta de IA, señalar un año aporta poca información. Localizar la afirmación, la fuente pertinente y la fecha de corte permite plantear una objeción comprobable.

Ocho decisiones que el lector puede reproducir

Las filas proceden de la evaluación publicada. Las **respuestas en inglés se construyeron para el experimento**. Conservamos sus identificadores, cortes y predicciones. **P** designa una afirmación que da por consumado el suceso cuando aún no había ocurrido en la fecha de corte, **O** indica otro caso y **R** significa revisión. La segunda columna resume en español el contenido del original. El corpus incluye los textos exactos.

Caso y corte	Qué hace la respuesta	Referencia	Año	Fechas
perseverance-04 · 2020-12-31	Prevé que el róver aterrice en 2021. El plan ya era público.	O	P	O
perseverance-05 · 2020-12-31	Niega que exista confirmación de un aterrizaje en 2021.	O	P	O
osiris-06 · 2022-12-31	Repite una pregunta acerca de una entrega en 2023.	O	P	O
rosetta-08 · 2013-12-31	Habla de dudas sobre el calendario y después afirma que Rosetta despertó.	P	O	P
dart-09 · 2022-01-01	Dice que DART impactó contra Dimorphos en 2022. El impacto fue en septiembre.	P	O	P
h03 · 2022-11-01	Imagina qué podrían hacer los científicos si la cápsula ya hubiese llegado.	O	O	O
h10 · 2014-12-31	Relata el despertar de Rosetta de enero de 2014 después del suceso.	O	O	R
h16 · 2021-02-10	Dice que el róver se diseñó para aterrizar y que «lo hizo» en febrero.	P	O	R

Una fecha futura puede proceder de la pregunta que la respuesta cita. Una reserva en la primera oración puede coexistir con una afirmación inequívoca en la siguiente. El detector puede pasar por alto una afirmación que recurre a un pronombre o interpretar una hipótesis como un hecho consumado. La exactitud factual merece otra valoración. En el piloto histórico, una respuesta situó en 2021 un relato

distinto del accidente aéreo documentado. Eso exige consultar las fuentes aunque el año no active ninguna alerta. También conviene contar por separado las peticiones fallidas, las respuestas que declinan contestar y los casos que el detector remite a una persona.

Qué fecha acredita cada fuente

Conviene registrar la fecha del suceso, la del anuncio o publicación, el momento en que un tercero observó una versión concreta y el corte fijado en la consulta. Una norma o contrato puede tener además una fecha de entrada en vigor. Un gobierno puede anunciar hoy el cambio del tipo de interés de mañana. En ese intervalo el cambio y su fecha prevista ya pueden ser públicos. Para juzgar la respuesta hay que saber si afirma que el gobierno anunció el cambio o que el nuevo tipo ya estaba en vigor. El acuerdo y el cierre de la operación bancaria del piloto presentan la misma dificultad.

En el corpus, el impacto de DART ocurrió el 26 de septiembre de 2022. La fuente de la NASA seleccionada para acreditarlo se publicó el día 27. Para una consulta entre ambos hitos cabe afirmar que esa fuente concreta aún no estaba disponible. Otras podrían haber dado la noticia antes. El estudio usa días enteros y fichas redactadas a partir de fuentes. La fecha que muestra hoy una página editable tampoco acredita por sí sola el texto que mostraba entonces. De ser posible, consérvense los bytes recuperados o una cita autorizada, el identificador de versión, la fecha de captura, la zona horaria y la huella del contenido. Una captura acredita cuándo ese archivo observó una versión. Podría haber estado disponible antes. Informe de la NASA · Acceso HTTP a versiones anteriores · Problemas de la evaluación temporal

Propuesta de adjudicación humana

1. **Fijar la pregunta y la proposición.** Registrar el corte y la zona horaria, conservar la pregunta y dividir la respuesta en afirmaciones verificables. Mantener el contexto suficiente para interpretar negaciones, citas, discurso referido y pronombres. Marcar qué datos aporta la propia pregunta. Registrar por separado el respaldo histórico y las posibles vías por las que llegó la información al modelo.
2. **Asignar dos evaluadores independientes a cada caso.** Darles el mismo conjunto de evidencias y ocultar las predicciones automáticas, la identidad del modelo y las etiquetas del otro evaluador. Para cada proposición, registrar qué expresa (hecho consumado, pronóstico, negación, pregunta o cita, abstención o ambigüedad), la fuente fechada y la relación temporal entre el suceso y el corte. Un calendario puede respaldar un pronóstico. Para afirmar que el acto previsto ocurrió se requiere evidencia de realización.
3. **Admitir una referencia indeterminada.** Las fuentes pueden discrepar, los archivos pueden tener lagunas y una frase puede referirse a actuaciones diferentes. Hay que identificar la información que falta y las interpretaciones que cambiarían el resultado. Los casos indeterminados siguen contando en la muestra. Las etiquetas binarias de los ejemplos publicados se fijaron al prepararlos y no se han sometido a doble anotación ciega.
4. **Medir el desacuerdo antes de resolverlo.** Publicar recuentos por evaluador, la matriz de desacuerdo y la concordancia observada para la interpretación del texto y para la fecha, por separado. Puede añadirse una medida de concordancia ajustada por azar, con sus supuestos y las frecuencias por clase. Una tercera persona puede resolver discrepancias explicando el motivo por escrito. Conservar ambos votos iniciales y la decisión final. Artstein y Poesio, concordancia entre anotadores

Cómo llevar la evaluación a un servicio real

En un estudio prospectivo, definir antes de conocer los resultados el servicio, la población, las lenguas, el periodo y la forma de seleccionar las preguntas. Muestrear peticiones ordinarias e incluir preguntas negativas, fechas explícitas, datos anticipados por el usuario y sucesos que nunca ocurren. Formar un panel de fuentes de referencia separado del corpus de recuperación. Congelar la primera captura y registrar las correcciones y ampliaciones conforme se observen. Repetir consultas antes y después de un anuncio, de la ejecución del acto y de una rectificación. Reservar juntas todas las versiones de cada suceso para que dos preguntas casi idénticas no queden a ambos lados de la evaluación. Añadir dominios, idiomas y sucesos nuevos. Una muestra enriquecida sirve para medir la respuesta a dificultades específicas. Para estimar con qué frecuencia las encuentra un usuario hace falta una muestra representativa del servicio. ExAnte

Publicar los totales de peticiones, fallos de recuperación y de proveedor, abstenciones del modelo, respuestas utilizables, afirmaciones elegibles, referencias indeterminadas y derivaciones a revisión. Para cada método, informar de la sensibilidad sobre **todos los positivos con referencia resuelta**, la tasa de alarma entre los negativos y la precisión de las alarmas emitidas cuando pueda calcularse. Dar también la cobertura (casos decididos automáticamente entre los elegibles), el error en esos casos y la proporción de positivos enviados a revisión. Si el servicio incorpora una cola humana, medir el tiempo de revisión y evaluar por separado el acierto y el retraso del proceso completo. Cuando haya suficientes sucesos independientes, calcular la incertidumbre remuestreando sucesos enteros. Las paráfrasis de cuatro misiones no equivalen a decenas de sucesos independientes. Clasificación selectiva y cobertura

El desafío congelado ilustra la importancia del denominador. El comparador con fechas decidió automáticamente dos de los 16 casos, ambos controles. Alertó sobre **0 de 8** positivos de referencia y derivó los ocho a revisión. La carga de revisión alcanzó **14 de 16**. No se midió el tiempo de una cola humana. Informar solo de las dos decisiones automáticas correctas ocultaría el trabajo pendiente para el auditor.

Qué puede preguntar la ciudadanía al amparo del Reglamento de IA de la UE

Quien impugna una decisión en la que intervino una IA puede necesitar saber qué versión de una norma utilizó el sistema. El estudio de SPE Audit registra la fecha de corte, los resúmenes de las fuentes seleccionadas, la petición exacta, una respuesta escrita para la prueba y el veredicto de la regla. Así se puede examinar la evidencia de una ejecución. Una huella digital permite detectar diferencias con una copia de referencia. Para saber cuándo estuvo disponible una fuente, qué norma estaba en vigor y en qué se basó la persona que resolvió hacen falta otras pruebas.

Quién debe poder explicar cada cosa

Artículo	Quién responde	Evidencia que puede aportar el estudio	Qué debe acreditar además la organización
12 · Registro automático	Proveedor del sistema de alto riesgo	El registro enlaza una ejecución con su fecha de corte, sus entradas y su salida.	El sistema puede registrar automáticamente sucesos pertinentes durante su vida útil, incluidos los necesarios para detectar riesgos y modificaciones sustanciales y vigilar su funcionamiento tras el despliegue.
13 · Instrucciones de uso	Proveedor del sistema de alto riesgo	Hay ejemplos de errores ante pronósticos, paráfrasis y fechas del mismo año.	El responsable del despliegue recibe instrucciones claras sobre finalidad, capacidades, limitaciones, supervisión e interpretación en condiciones de uso reales.
14 y 26.2, 26.5 · Supervisión humana	El proveedor diseña la supervisión. El responsable del despliegue designa personas y vigila el funcionamiento.	Se pueden leer juntos la respuesta y los documentos admitidos por la fecha de corte.	El personal designado tiene competencia, formación, autoridad y apoyo para intervenir. El expediente recoge qué hizo ante una alerta o un incidente concreto.
19 y 26.6 · Conservación	Proveedor y responsable del despliegue, cada uno por los registros bajo su control	El formato permite conservar y volver a comprobar un registro.	Cada uno conserva los registros automáticos durante un tiempo adecuado a su finalidad, al menos seis meses salvo que el Derecho de la UE o nacional aplicable disponga otra cosa. Debe gestionar el acceso y la supresión.
50.1, 50.2 y 50.4 · Transparencia	Proveedores de interfaces y generadores sujetos a la norma. Responsables del despliegue que publican determinados contenidos.	Queda constancia de la salida que puede inspeccionarse.	Según el uso y sus excepciones, procede avisar de la interacción con IA, marcar contenidos sintéticos en formato legible por máquina o comunicar su origen al publicarlos. Hace falta verificar cómo se muestra el aviso y su accesibilidad.
85 y 86 · Reclamación y explicación	La autoridad de vigilancia del mercado recibe reclamaciones. El responsable del despliegue explica las decisiones cubiertas.	Una discrepancia de fechas permite formular una pregunta concreta.	La autoridad tramita la presunta infracción según su procedimiento. Para explicar el papel de la IA y los principales elementos de la decisión hace falta el expediente.

El Reglamento sí prescribe campos mínimos para la identificación biométrica remota del anexo III, punto 1.a. El artículo 12.3 incluye el período de uso, la base de datos de referencia, las entradas que produjeron una coincidencia y la identidad de quienes verificaron los resultados. Los campos del estudio responden a su pregunta de investigación. En un servicio real, el tratamiento de datos personales exige base jurídica, minimización, seguridad adecuada y límites de conservación. Se puede publicar un ejemplo sintético mientras el personal autorizado examina el expediente individual.

Cuándo se aplica

- **2 de agosto de 2026.** Fecha general del artículo 50. Según el artículo 111.4, los proveedores de sistemas generadores de contenido sintético sujetos a la norma que ya se comercializaban antes de esa fecha tienen hasta el **2 de diciembre de 2026** para cumplir el marcado del artículo 50.2.

- **2 de diciembre de 2027.** Las secciones 1 a 3 del capítulo III, salvo el artículo 6.5, se aplican a los sistemas de alto riesgo del artículo 6.2 y del anexo III.
- **2 de agosto de 2028.** Esas secciones se aplican a los sistemas de alto riesgo vinculados a productos del artículo 6.1 y del anexo I. El artículo 111.2 contiene reglas transitorias para sistemas anteriores, con un plazo hasta 2030 para los usos por autoridades públicas comprendidos en él.

Ejemplo hipotético de reclamación por una prestación pública

Supongamos que un ayuntamiento pone en servicio en 2028 un sistema de alto riesgo para evaluar el acceso a una prestación asistencial pública esencial, un uso del anexo III, punto 5.a. Deniega una solicitud. La persona afectada observa que la motivación cita una reforma publicada después de la fecha de referencia. Puede preguntar al ayuntamiento qué versión debía aplicarse, qué aportó la IA y por qué la persona que resolvió se basó en ella. El personal puede contrastar la cápsula con las fechas oficiales de publicación y entrada en vigor, la solicitud, la resolución firmada y la actuación de quien la revisó. La publicación posterior justifica investigar. El posible efecto jurídico de la reforma y el resultado para la persona requieren examinar la norma y su expediente.

El artículo 86 cubre determinadas decisiones basadas en resultados de sistemas de alto riesgo del anexo III, salvo su punto 2, con efectos jurídicos o de importancia similar que la persona considere perjudiciales para su salud, seguridad o derechos fundamentales. Pueden aplicarse excepciones o restricciones legales. El artículo opera en la medida en que el Derecho de la UE no reconozca ya ese derecho. Cualquier persona física o jurídica con motivos para considerar infringido el Reglamento puede reclamar ante la autoridad de vigilancia del mercado competente según el artículo 85. La impugnación de la propia denegación de la ayuda sigue la vía administrativa o judicial que corresponda.

Fuentes oficiales consultadas el 21 de noviembre de 2026. Reglamento de IA, consolidado a 27 de julio de 2026 · Reglamento modificativo (UE) 2026/1744 · Comisión, anexo III · Derecho a presentar una reclamación · Derecho a obtener explicaciones · RGPD, artículos 5 y 6.

Cómo se sitúa este estudio entre las revisiones publicadas

Selección. Realizamos búsquedas dirigidas en inglés hasta el 21 de noviembre de 2026. Combinamos *survey* o *review* con consultas sobre preguntas y recuperación temporales, evaluación de sistemas RAG, verificación de hechos, alucinaciones, contaminación de datos, procedencia digital y filtraciones en pronósticos. Seleccionamos fuentes que tratan distintas preguntas de medición de este estudio y cuya fecha y alcance se podían comprobar en una editorial, congreso, organismo público o prepublicación de los autores. La tabla incluye cinco revisiones, un informe técnico oficial y un análisis metodológico. Las fechas identifican la versión citada. La selección es dirigida y carece de un recuento exhaustivo de publicaciones o citas.

Obra y fecha	Qué estudia	Cómo orienta la auditoría	Acceso
Guo, Schlichtkrull y Vlachos, *A Survey on Automated Fact-Checking* (TACL, 2022)	Selección de afirmaciones, recuperación de pruebas y verificación. Compara definiciones y conjuntos de datos.	Conviene anotar exactamente qué afirma una frase. Una pregunta que cita un desenlace, una negación y una afirmación de que sucedió merecen etiquetas distintas.	Artículo y PDF

Obra y fecha	Qué estudia	Cómo orienta la auditoría	Acceso
Huang et al., *A Survey on Hallucination in Large Language Models* (prepublicación, nov 2023)	Distingue los errores de veracidad y de fidelidad al contexto. Revisa causas, detección y evaluación.	Un hecho verdadero hoy pudo ser desconocido en la fecha consultada. Su veracidad y su respaldo histórico requieren campos distintos.	Manuscrito abierto · Revista
Chandra et al., *Reducing Risks Posed by Synthetic Content* (NIST AI 100-4, nov 2024)	Informe técnico oficial sobre procedencia, autenticación, etiquetado, marcas de agua y detección.	Una huella conservada permite comparar unos bytes concretos. La autenticidad de la fuente, la integridad de la captura y la veracidad de una frase exigen pruebas específicas.	Informe y PDF
Cheng, Chang y Wu, *A Survey on Data Contamination for Large Language Models* (prepublicación, feb 2025)	Solapamiento entre entrenamiento y evaluación, prevención y detección según el acceso disponible al modelo.	Un hecho posterior en una respuesta tiene varios orígenes posibles, como la pregunta, las fuentes recuperadas o el entrenamiento. La respuesta sola no identifica el origen.	Manuscrito abierto
Gan et al., *Retrieval Augmented Generation Evaluation in the Era of Large Language Models* (prepublicación, abr 2025)	Revisa cómo se evalúan la recuperación, la generación, el sistema completo, la exactitud, la seguridad y la eficiencia.	Revisar qué documentos entraron y revisar cómo se clasificó la respuesta resuelve preguntas diferentes. Registrar ambas operaciones permite localizar el error.	Manuscrito abierto
Paleka et al., *Pitfalls in Evaluating Language Model Forecasters* (prepublicación de mayo de 2025, ICLR 2026)	Análisis metodológico de información futura en evaluaciones retrospectivas. Examina artículos actualizados y efectos de la recuperación.	Un documento fechado antes del corte puede contener una revisión posterior. Una afirmación histórica exige conservar la versión o declarar qué se supone sobre ella.	Prepublicación · Congreso
Piryani et al., *It's High Time: A Survey of Temporal Question Answering* (ACL, jul 2026)	Ordena las tareas según el tiempo del corpus, el de la pregunta y las capacidades del modelo. Revisa RAG temporal, expresiones de tiempo y evaluación.	La fecha que figura en la pregunta, la disponibilidad del documento y el momento del suceso pueden diferir. El protocolo debe decir cuál determina la admisión.	Artículo y PDF

Un antecedente experimental directo. Allein, Augenstein y Moens estudiaron cómo las fechas de las afirmaciones y de sus pruebas afectan al orden de recuperación y al veredicto. Su resultado justifica registrar ambas fechas. Saber si un proveedor procesó los documentos elegidos requiere verificar la ejecución concreta. Journal of Web Semantics, nov 2021 · Manuscrito abierto.

La aportación que se defiende aquí es una inspección reproducible de una implementación pública identificada, con una decisión documentada para cada caso. Las revisiones aportan terminología, fallos conocidos y puntos de comparación. Una afirmación de prioridad o exhaustividad en todo el campo requeriría otro diseño de búsqueda.

Revisión de noviembre. Qué se sabía sobre auditoría temporal de la IA

Anexo de literatura y documentos de patente · 21 de noviembre de 2026 · Edición española

Demostrar que una respuesta existía, identificar qué documentos se seleccionaron y atribuir una respuesta al uso de información futura son tres problemas con requisitos probatorios distintos. Sistemas anteriores

ya abordaban partes importantes de los tres. Esta revisión identifica los documentos más próximos que hemos podido comprobar, indica dónde aparece cada mecanismo y explica su relevancia para los experimentos. Para la descripción técnica original de SPE se toma como referencia histórica el 4 de diciembre de 2025. El software público tiene su propia fecha y la investigación científica incorpora también trabajos posteriores.

Cómo se hicieron las comparaciones

Se siguieron referencias y se buscaron publicaciones científicas y documentos de patente sobre recuperación histórica, fugas de información futura, procedencia del contexto, inferencia verificable, confianza semántica y contraste con bases externas. Se comprobaron también fechas anteriores de preprints y de solicitudes publicadas posteriormente. Cada aportación se asignó al objeto sobre el que opera: datos de entrenamiento, índice de recuperación, registros seleccionados, carga enviada al modelo, afirmaciones generadas o corpus de referencia del evaluador. Que dos trabajos utilicen la expresión «prueba», «confianza» o «tiempo» no basta para establecer equivalencia.

La numeración que sigue resume en palabras propias dos generaciones de diseño. Permite leer este anexo sin acceso a archivos privados. Los números 1–13 proceden de la descripción original y los números 14–27, de borradores de mayo de 2026. Ninguno implica que se haya concedido o interpretado oficialmente una reivindicación de SPE. Para los documentos de terceros se cita la publicación que describe el mecanismo. Cuando se analiza una reivindicación concedida se identifica la concesión por separado. El trabajo es una búsqueda dirigida de antecedentes y carece de alcance mundial exhaustivo o de dictamen sobre libertad de operación.

Documentos más próximos al diseño

N01. Dantart. Recuperación histórica con un registro firmado de cada consulta

La primera versión de la arquitectura forense de Alex Dantart se depositó en arXiv el **12 de octubre de 2025**. La sección 3.2 y la figura 1 describen una ingestión con versiones, huellas criptográficas de documentos, particiones según la vigencia jurídica, aplicación de políticas por un orquestador y escritura de un registro JSON firmado en almacenamiento de una sola escritura. El registro puede contener la consulta, las huellas de los documentos recuperados, el contexto temporal, la respuesta y la evaluación de la política. Las secciones 4.4, 5.1 y 5.3 desarrollan el protocolo experimental, la medida de validez temporal y un ejemplo del registro. La fecha de publicación de una norma, su entrada en vigor y el momento en que un operador obtuvo su copia pueden ser distintos. Hay que declarar cuál gobierna la reconstrucción.

Es el precedente más directo para reunir recuperación histórica y registro resistente a alteraciones en una misma consulta. En los pasajes leídos no aparecen conjuntamente la puntuación semántica concreta de tres componentes, un compromiso independiente sobre la referencia externa y el uso combinado de ambos para clasificar la respuesta. El artículo sitúa su evaluación en un entorno sintético. Su marco jurídico o sus estimaciones de rendimiento no se trasladan directamente a otro sistema. La revisión del **23 de diciembre** es posterior al corte. Los mecanismos anteriores ya figuran en octubre. Texto de octubre, §§3.2, 4.4, 5.1 y 5.3 · Historial y DOI 10.48550/arXiv.2510.12830

N02. EQTY Lab. Evidencia criptográfica sobre la entrada utilizada en la inferencia

Verifiable computing, **US 2025/0254026 A1**, reivindica fecha de prioridad de 1 de febrero de 2024. La solicitud estadounidense se presentó el 31 de enero de 2025 y se publicó el **7 de agosto de 2025**. En los ejemplos consecutivos sobre RAG se describen huellas de fuentes, representaciones vectoriales y metadatos autenticados, una raíz asociada a una versión del corpus y la comprobación de si un elemento recuperado se incluyó en la entrada del modelo. La reivindicación 1 publicada trata la generación de una clave de atestación en un entorno confiable, la declaración sobre sus entradas y salidas, la firma y la exportación. Los ejemplos de recuperación se encuentran en la descripción. No cabe asignarles automáticamente el alcance de una reivindicación concedida.

El documento va más allá de comprobar la presencia de un registro en el *entrenamiento*. Quien estudie la exclusión en *inferencia* debe contrastar este ejemplo. Comprobar la inclusión de un elemento concreto y repetir una política temporal sobre el conjunto identificado de candidatos son operaciones diferentes. La atribución al proveedor de los bytes realmente recibidos depende asimismo del límite de la atestación. La solicitud posterior **US 2026/0230327 A1**, presentada el 31 de marzo y publicada el 6 de agosto de 2026, amplía el contexto a varios agentes y cita la familia anterior. La pertenencia a la familia no fecha todo el contenido nuevo en 2024. Solicitud estadounidense y reivindicación 1 · Publicación internacional WO 2025/166192 A1 · Desarrollo posterior

N03. Liu y colaboradores. Evaluar si una respuesta afirma un hecho verificable solo más tarde

ExAnte v1, con fecha **26 de mayo de 2025**, define conjuntos de hechos de referencia anteriores y posteriores a un corte. La sección 2.1 evalúa desenlaces únicos y afirmaciones atómicas. La 3.3 construye pares de referencias a partir de versiones históricas de Wikipedia. La 3.4 compara títulos de publicaciones con su primera fecha pública conocida. Así se puede distinguir lo que afirma un texto de lo que se podía verificar en esa fecha. La posterior publicación en EACL 2026 no cambia la fecha del preprint.

Para una alerta de violación temporal en la salida, este antecedente es más cercano que la investigación genérica sobre alucinaciones. También revela una decisión de medición. Un conjunto de evaluación puede clasificar una predicción acertada como fuga por la definición de su tarea. Para distinguir previsión y afirmación retrospectiva hace falta una regla de anotación explícita. La coincidencia con el desenlace o con una referencia posterior no identifica por sí sola la causa interna de la generación. El procedimiento revisado tampoco vincula mediante un compromiso criptográfico esa clasificación con los bytes de la entrada exacta. Preprint de mayo, §§2.1, 3.3 y 3.4 · EACL, DOI 10.18653/v1/2026.eacl-long.72

N04. Bitvore. Observaciones externas fechadas y puntuación de alucinación

US 2025/0086211 A1, presentada el 12 de septiembre de 2023 y publicada el **13 de marzo de 2025**, extrae métricas de la comparación de una salida con una base de referencia. En el párrafo [0082], un ejemplo de relaciones empresariales incorpora entidades, relevancia, confianza en la extracción, número de registros que respaldan la relación y fecha desde la que hay evidencia. En la descripción se analiza también la exactitud temporal. La reivindicación 1 publicada une consulta, generación, contraste según varios criterios y cálculo de una puntuación. Cuestiona la novedad de la operación genérica de comparar la respuesta con señales externas fechadas y marcar escasez de respaldo. Queda abierta la comprobación de qué registros llegaron al modelo en una ejecución específica.

La USPTO **concedió US 12,783,564 B2 el 3 de noviembre de 2026**. La reivindicación 1 concedida añade condiciones sobre fuentes en tiempo real y estructuradas, un conjunto de modelos y la modificación iterativa y selectiva de la consulta. El alcance de esta reivindicación no puede describirse copiando la redacción más amplia de la solicitud de marzo de 2025. La distinción es material porque una ficha bibliográfica consultada seguía mostrando el expediente como pendiente. Solicitud publicada · PDF del documento, ¶0082 · Gaceta oficial USPTO, concesión y primera reivindicación

N05. Qiu y Miikkulainen. Densidad semántica para medir confianza

El artículo de NeurIPS **2024** define una confianza por respuesta en el espacio semántico. La sección 3.3 explica cómo estimar la densidad y la 4 presenta la evaluación con tareas de preguntas y respuestas. Ya se utilizaba, por tanto, una densidad en ese espacio para estimar incertidumbre. La estadística de los autores difiere de la propuesta de SPE de combinar coherencia entre ejecuciones, estabilidad ante perturbaciones y densidad de reconocimiento con pesos definidos. Para valorar la diferencia conviene mantener fijos el corte y las fuentes, variar los componentes, y medir si mejoran la detección de afirmaciones sin respaldo. La coincidencia terminológica no proporciona ese resultado. Actas, DOI 10.52202/079017-4274 · Artículo, §§3.3 y 4

N06. Paleka y colaboradores. Contenido futuro dentro de documentos aparentemente admisibles

El preprint *Pitfalls in Evaluating Language Model Forecasters* apareció el **31 de mayo de 2025**. La sección 2.2 describe páginas antiguas actualizadas después con el desenlace y sesgos de ordenación que sobreviven al filtro de fecha. Examina el filtrado adicional por el contenido de los artículos recuperados, el recurso a fuentes cuya fecha es más fiable y el mantenimiento de índices históricos. También señala un sesgo sutil: eliminar una página actualizada puede suprimir información contradictoria que sí habría leído el pronosticador en la fecha histórica. La sección 2.3 trata la incertidumbre sobre los cortes de los propios modelos.

Este trabajo identifica el problema técnico y aporta una motivación documentada para examinar el contenido antes de admitir un registro. Afecta en especial a los borradores 14–15 de SPE, que cuentan referentes temporalmente sospechosos y prevén una política para los documentos sin referentes extraídos. Los autores no especifican los umbrales concretos, la política alternativa ni los campos de la cápsula propuestos después. Una referencia rara puede ser genuina. Por ello, los controles deberían incluir términos históricos poco frecuentes y anuncios públicos de planes para el futuro. Preprint, §§2.2 y 2.3 · Fecha e identificadores

N07. Intuit. Perturbación de la consulta y acuerdo entre modelos

Intuit presentó la solicitud estadounidense el 30 de agosto de 2023. Se publicó como **US 2025/0077777 A1 el 6 de marzo de 2025** y se concedió como **US 12,387,049 B2 el 12 de agosto de 2025**. La reivindicación 1 concedida genera formulaciones semánticamente equivalentes, compara respuestas de dos modelos y calcula una medida de consistencia. La 20 detalla una disposición de muestras. La atribución a IBM que figuraba en una revisión anterior es errónea. La solicitante original es Intuit.

Existe solapamiento directo con parte de la coherencia y de la estabilidad ante perturbaciones. El pasaje no establece la combinación concreta de tres medidas, las representaciones acotadas por fecha y una referencia externa versionada. Dos modelos pueden coincidir por compartir material de entrenamiento o un mismo error. Un conjunto de evaluación temporal debería contener ejemplos en los que varios modelos puedan repetir un error histórico plausible. Solicitud publicada · Concesión, reivindicaciones 1 y 20

N08. IBM. Inclusión en un conjunto de entrenamiento

US 2025/0156390 A1 de IBM, presentada el 14 de noviembre de 2023 y publicada el **15 de mayo de 2025**, organiza un conjunto de entrenamiento en un árbol de Merkle y genera una prueba referida a un elemento y su raíz. La reivindicación 1 incluye entrenar un LLM con ese conjunto. La concesión **US 12,608,358 B2** se publicó el **21 de abril de 2026**. La pertenencia al conjunto de entrenamiento difiere de la pertenencia a la entrada de una consulta. Un modelo puede recibir un documento para responder aunque la prueba sobre el entrenamiento no lo mencione. También puede conservar en sus parámetros información que no apareció en la recuperación. IBM es un antecedente para estructuras autenticadas. El ejemplo de inferencia de EQTY aborda de manera más directa la cuestión del contexto admitido. Solicitud, reivindicación 1 · Gaceta oficial USPTO

N09. Vodafone. Publicación después del corte con solicitud anterior

US 2026/0017491 A1 indica prioridad británica de 9 de julio de 2024, presentación estadounidense de 2 de julio de 2025 y publicación de **15 de enero de 2026**. Su primera reivindicación codifica el vector de salida de un modelo RAG de dominio cerrado con un autocodificador variacional entrenado con ejemplos etiquetados como normales o alucinatorios. La distribución aprendida sirve para identificar alucinaciones candidatas. Es una interpretación alternativa de una anomalía semántica. No acredita una exclusión temporal de fuentes ni verifica el resumen criptográfico de una entrada concreta. Además de su importancia científica, ilustra una distinción jurídica: bajo las condiciones legales pertinentes, un documento estadounidense publicado más tarde puede ser oponible desde una fecha efectiva de presentación anterior. Para ello hay que demostrar el soporte de la materia que se opone. Publicación y familia · Manual USPTO §2154

N10. Shukla y Joshi. Recibos de evidencia por afirmación

Proof-Carrying Answers vincula las afirmaciones de una respuesta con la evidencia recuperada mediante, entre otros elementos, información autenticada del índice vectorial y recibos criptográficos. El programa de ACSAC contiene una presentación fechada el **8 de diciembre de 2025**, cuatro días después de la fecha de comparación. Esta revisión no identificó una versión pública anterior al corte con fecha comprobada. La granularidad es importante: la cápsula de una salida completa y el recibo de una afirmación responden a preguntas diferentes. Un pasaje que sustenta una frase puede estar desactualizado o usar una convención temporal distinta de la que presupone el lector. Es un comparador valioso para diseñar el informe aun cuando no está acreditado aquí como divulgación pública anterior. Taller y presentación fechada · Actas IEEE, DOI 10.1109/ACSACW69556.2025.00050

Correspondencia por rasgo técnico

Rasgo del diseño de SPE	Enseñanza más próxima comprobada	Qué queda por establecer
1(a), fecha objetivo y política admisible	N01 usa la fecha de consulta, vigencia de las normas y política en el orquestador.	Qué convención temporal gobierna y cómo se demuestra la fiabilidad del dato de fecha.
1(b), intercepción, exclusión y recuperación acotada	N01 particiona el corpus jurídico por fecha. N06 muestra contenido actualizado que sobrevive a una restricción nominal.	Deben poder reconstruirse los candidatos, la entrada real, la política y los elementos excluidos o sin determinar.
1(c), compromiso determinista sobre elementos ordenados	N01 firma una traza con huellas de los documentos. N02 describe recuperación autenticada e inclusión en la entrada. RFC 8785 aporta un precedente de JSON canónico.	El resumen vincula el manifiesto. Para demostrar integridad respecto a todo el índice y entrega al proveedor hacen falta pruebas adicionales.
1(d), ejecución del modelo y captura de su salida	N01 registra la respuesta. N02 declara y firma entradas y salidas.	La captura en el lado cliente necesita un límite explícito de confianza antes de poder atribuir la entrada efectiva al modelo remoto.
1(e), medida de coherencia, estabilidad y reconocimiento	N05 estima densidad. N07 mide consistencia bajo formulaciones equivalentes entre modelos.	Un único pasaje de los citados no establece la puntuación exacta de tres componentes dentro de la cadena completa. Se requiere medir el aporte de cada componente conservando los mismos datos.
1(f), contraste con una referencia externa fechada	N03 construye referencias históricas. N04 usa respaldo, fechas y puntuaciones.	Comprobar la disponibilidad del corpus, documentar normalización y valores ausentes, y mantener fuera de la entrada las fuentes posteriores reservadas al evaluador.
1(g), clasificación de la integridad semántica	N03 evalúa cuándo se podía verificar una afirmación. N04 y N07 miden alucinación o consistencia con evidencias diferentes.	Declarar reglas para planes, negaciones, hipótesis, citas, hechos falsos, fuentes faltantes y casos sin resolver.
2, cápsula con identificadores, huellas, tiempo y política	N01 conserva registros por consulta. N02 atesta entradas y salidas.	Identificar quién crea cada campo y quién puede autenticar de forma independiente fecha, firmante, instantánea y vínculo con la respuesta.
4(i), afirmación sobre un hecho posterior con contexto acotado	N03 evalúa afirmaciones dependientes de información futura. N01 y N02 aportan evidencia de recuperación de distinto alcance.	La mención de un año posterior y la afirmación de un desenlace son observables distintos. La causa interna de la generación queda latente.
4(ii), puntuación interna alta con poco respaldo externo	N04 evalúa respaldo escaso. N05 y N07 miden confianza y consistencia.	Un archivo incompleto, una lengua infrarrepresentada o un índice ausente pueden producir la misma discrepancia.
9, representación temporal e informe de divergencia	N03 aborda la referencia temporal. N05 y N09 describen confianza o distribuciones latentes.	La redacción de sistema deja por demostrar qué fallo verificable del mundo real sigue la divergencia calculada.
Borradores 14–15, examinar contenido recuperado y tratar documentos sin referentes	N06 motiva explícitamente el filtrado posterior a la recuperación. RFC 3161 aporta sellado de tiempo.	El estadístico de cuarentena y la política alternativa concretos se desarrollan después. Una fecha autenticada no garantiza la admisibilidad semántica.
Borrador 26, repetir la política sobre un índice identificado y recomputar el compromiso	N01 registra la consulta. N02 describe explícitamente inclusión o exclusión durante la inferencia.	Fijar versiones, parámetros de búsqueda, orden de candidatos, desempates, política y bytes entregados. El borrador 26 también depende del método completo de la reivindicación 1.

Canonización JSON, RFC 8785 · Sellado de tiempo, RFC 3161 · Inclusión y consistencia en árboles de Merkle, RFC 9162

Novedad, fechas y valor experimental

Una fuente y una combinación de fuentes. No se identificó en el conjunto revisado una divulgación anterior al corte que contenga todos los elementos 1(a–g) en la disposición reivindicada. Sí existe un argumento concreto de obviedad. Quien guarda fuentes históricas y una respuesta firmada (N01) tiene un motivo para averiguar si el modelo cuenta hechos posteriores (N03). N02 ofrece una técnica de procedencia de la entrada que un tercero podría comprobar. N04, N05 y N07 ofrecen antecedentes del contraste externo, la confianza o la consistencia. El análisis de patentes aún exigiría justificar exactamente qué combinaría una persona experta, por qué y con qué expectativa de éxito. La posible ventaja de la combinación exacta sigue siendo una pregunta experimental. De este examen no se deduce que se infrinja una reivindicación de un tercero o que sea válida o inválida. USPTO, anticipación · USPTO, obviedad

Detalles posteriores. El texto original contempla listas acotadas, huellas, instantáneas del índice y parámetros versionados. El borrador 26 precisa la repetición del proceso por un tercero. Los borradores 14–15 aplican análisis semántico durante la recuperación y establecen una política para los documentos sin referentes extraídos. Otros borradores posteriores describen comprobaciones del entorno de ejecución, el truncamiento y las versiones del modelo. Asignar una fecha anterior exige rastrear el soporte de cada rasgo. En la revisión también apareció **US 2025/0378436 A1**, publicada el **11 de diciembre de 2025**. Su objeto son peajes con moneda digital. La canonización de los mensajes de transacción ofrece una analogía técnica distante. No constituye aquí una publicación anterior acreditada sobre auditoría RAG. USPTO, beneficio de la fecha anterior · Publicación sobre peajes, descripción de la transacción 114

Controles derivados de trabajos independientes. SelfCheckGPT analiza el desacuerdo entre muestras sin usar una base factual externa. Farquhar y colaboradores miden mediante entropía semántica la incertidumbre asociada a confabulaciones. Halawi y colaboradores acotan la recuperación a una fecha histórica simulada en pronósticos. El Lahib y colaboradores, cuyo preprint apareció el **31 de enero de 2026**, muestran cómo un buscador puede entregar contenido actualizado pese a una restricción de fecha. De estos trabajos se desprenden controles posibles: salidas idénticas con contextos distintos, salidas semánticamente distintas con el mismo año, previsiones plausibles, negaciones, citas, hechos verificables solo posteriormente y afirmaciones sin resolver. Conviene conservar por separado la instantánea de recuperación y las fuentes de adjudicación, e informar de las abstenciones. Una declaración del proveedor sobre la entrada debe identificarse como tal mientras carezca de corroboración independiente. SelfCheckGPT, EMNLP 2023 · Entropía semántica, Nature 2024 · Halawi et al., §3.2 · El Lahib et al., ACL 2026

Conclusiones

El hallazgo más fácil de reproducir es un fallo. Un vocabulario que funciona con la frase usada en el diseño pierde los ocho positivos de un pequeño desafío cuando el redactor expresa los mismos desenlaces de otras maneras. La utilidad operativa del comparador depende de quién atiende las abstenciones y con qué documentación. Un servicio que convirtiera «revisión» en un visto bueno ocultaría exactamente los casos cuya interpretación motivó el ensayo.

También hay una capacidad que se pudo demostrar con éxito. Un tercero puede conservar una huella, abrir la captura de la petición, reconstruir las fichas que pasaron el filtro, cotejar la respuesta registrada y encontrar una modificación. En las intervenciones puede además observar que el programa local eligió otra respuesta al retirar una ficha. Estas operaciones dan a dos personas que discrepan un objeto concreto

sobre el que discutir. Para responder quién conocía qué y cuándo, necesitan fuentes históricas, una atribución fiable y observaciones adicionales del entorno.

La revisión de antecedentes cambia la forma de presentar la contribución. La recuperación con fecha, los registros firmados, la procedencia, la evaluación del conocimiento futuro y el estudio de la incertidumbre tienen historias técnicas propias. Publicar las implementaciones estudiadas, sus límites y el resultado desfavorable permite que otro equipo cuestione el diseño y compare su alternativa sobre los mismos ejemplos.

Material, autoría y condiciones

Edición de investigación de **Rubén Abella · SPE Audit** El equipo que publica el estudio desarrolla la tecnología evaluada. La recuperación bibliográfica, la redacción, el código y la edición de los ejemplos recibieron asistencia de IA. El desafío fue redactado por separado por un editor asistido por IA, con conocimiento de la pregunta científica y sin acceso previo a las reglas específicas ni a los ejemplos de desarrollo. Las cifras se comprobaron regenerando los archivos y recontando las decisiones. Se trata de una primera edición pública sin revisión por pares ni evaluación independiente de despliegue. Las dos ediciones conservan los títulos de las fuentes en su lengua original. Las frases ensayadas están en inglés y las traducciones de la interfaz son editoriales.

Las fichas del experimento son resúmenes elaborados para este trabajo. Conservar la fecha de publicación que declara una agencia permite reproducir el filtro utilizado. Para una controversia sobre lo que una página decía en una fecha determinada haría falta una captura histórica identificada. Los extractos del piloto permiten recontar sus resultados preservando los turnos que fallaron. Los hashes de texto pueden cotejarse con el original que posea otra parte.

Reproducción. Descargue el paquete autónomo, extráigalo y ejecute `python3 verify.py` desde su carpeta con Python 3.10 o posterior. En Windows con el lanzador de Python, use `py -3 verify.py`. Abra `demo.html` para examinar casos en el navegador. La comprobación funciona sin claves, pago ni servicios externos. El informe de verificación identifica la plataforma efectivamente ensayada. La licencia del código y los términos aplicables se conservan en el repositorio público. Quien encuentre una fuente mal fechada o una afirmación mal clasificada puede abrir una incidencia pública indicando el identificador M, N, B o de caso y la evidencia concreta. Registro público de correcciones

Los 54 registros de la comparación de mayo

El índice original contenía 19 filas de patentes y 35 de artículos, software, protocolos, especificaciones y otras pistas. Cada registro conserva su posición de origen entre M01 y M54. Cuando una fila mencionaba varias publicaciones de patente o versiones de un manuscrito, se han mantenido juntas. El total de 54 corresponde a este índice. Los estudios y referencias adicionales tienen sus propios recuentos.

La fecha de comparación es el 4 de diciembre de 2025. Hemos recuperado 48 registros con fecha documental anterior y cuatro con fecha posterior. Las otras dos entradas plantean dificultades diferentes. ALETHEIA tiene commits fechados, pero carecemos de una prueba independiente de su primera fecha de acceso público (M37). No se ha conseguido identificar de forma inequívoca un módulo de Rust (M38). La clasificación se refiere a la edición o publicación citada. En arXiv se registra el primer envío, en una revista la publicación enlazada y en patentes la fecha del número de publicación indicado. Un repositorio actual o un documento de la misma familia pueden tener fechas distintas. Las versiones fechadas del software acreditan ejemplos concretos, sin establecer cuándo apareció por primera vez la tecnología.

Qué puede comprobar una persona que audita

Pregunta	Registros pertinentes	Interpretación
¿Pertenecía el elemento seleccionado a la colección comprometida?	M01, M14–M16, M48	Se puede verificar la inclusión frente a la raíz de esa colección. Aún hay que identificar quién la reunió y qué pudo quedar fuera.
¿Ejecutó un programa la evaluación especificada sobre las entradas vinculadas?	M07, M23–M29, M31	Una prueba o atestación vincula un cálculo o entorno concretos. Hay que conocer sus entradas, salidas y controles. Las fechas que se suministran como datos de confianza siguen siendo supuestos.
¿Eran pertinentes los documentos para la fecha elegida?	M04–M05, M32–M36, M39–M43, M54	La pertinencia temporal, la vigencia jurídica y la disponibilidad pública en una fecha son propiedades diferentes. Conviene especificar qué propiedad se evalúa y con qué fecha se decide cada registro.
¿Respalda la evidencia suministrada una afirmación concreta?	M06, M09–M10, M44–M47	La consistencia lógica, el acuerdo entre respuestas, la atención al contexto y el respaldo en una fuente miden propiedades distintas. Una previsión citada o repetida debe diferenciarse de una afirmación sobre el desenlace.
¿Puede otra persona reconstruir la interacción registrada?	M02–M03, M08, M11–M15, M20–M22, M30–M31, M37, M48–M50, M54	Cada recibo cubre un objeto preciso. Los archivos de origen, la pregunta final tras el truncamiento, la versión del modelo y los resultados de herramientas pueden diferir aun cuando exista una huella válida en una fase anterior. La serialización determinista y la repetibilidad numérica corresponden a otras etapas.

La arquitectura más cercana. La v1 de Dantart, publicada en octubre de 2025, ya describe recuperación en índices temporales de fuentes jurídicas, evaluación de políticas y un registro JSON firmado de escritura única que vincula pregunta y respuesta con contexto temporal y huellas de documentos. Los apartados 3.2 y 5.3 contienen estos elementos en la versión de octubre. Una norma puede entrar en vigor después de su publicación o modificarse con efectos retroactivos. Un estudio que pregunte qué se podía saber públicamente necesita una fecha de disponibilidad, una versión de la fuente y una regla para cada afirmación. De esa diferencia surge una pregunta experimental concreta. Dantart v1, apartados 3.2 y 5.3 · Cronología de versiones.

Correcciones que cambian la lectura. La patente sobre consistencia semántica corresponde a Intuit (M09). La solicitud sobre peajes (M17) es periférica y su publicación A1 es posterior al corte, aunque el depósito estadounidense sea anterior. CausalRAG, CDF-RAG, KG-IRAG y Zep tienen manuscritos de principios de 2025 (M39, M40, M42 y M43). El registro distingue también un primer envío de una revisión (M34), un simposio de la conferencia principal (M35), un commit de su acceso público (M37) y la presentación de un producto de la comprobación independiente de su funcionamiento (M30). Los registros de patentes resumen divulgaciones técnicas concretas. Una conclusión sobre validez, alcance de las reivindicaciones o libertad de operación exigiría otro estudio.

Registro comentado de fuentes

Cada entrada fechada indica qué hecho acredita la fecha y su posición respecto al 4 de diciembre de 2025. Los enlaces relacionados identifican otras publicaciones o versiones cuando la fila original agrupaba documentos. El único enlace de M38 lleva al candidato identificado como tal.

Patentes ordenadas por el índice de revisión

M01 · Merkle-enhanced training data audit process. US20250156390A1 · International Business Machines Corp. 2025-05-15 · Publicación de patente · Anterior al corte.

Una prueba de Merkle vincula un elemento de entrenamiento al conjunto comprometido. En una consulta histórica también es necesario identificar los documentos entregados en esa ejecución. Registro principal · US12608358B2 · 2026-04-21 · USPTO · 2026-04-21

M02 · Providing explainable machine learning model results using distributed ledgers. US20220198304A1 · International Business Machines Corp. 2022-06-23 · Publicación de patente · Anterior al corte.

El registro conserva información para explicar un resultado de aprendizaje automático. Reconstruir esa procedencia deja abierta la admisibilidad temporal de los documentos recuperados. Registro principal

M03 · Producing and verifying computational determinations using a distributed ledger. US11303454B2 · International Business Machines Corp. 2022-04-12 · Publicación de patente · Anterior al corte.

Registra entradas y decisiones de un modelo y permite contrastarlas con el historial. La coincidencia de registros no verifica que una afirmación respete el corte histórico. Registro principal

M04 · Database record lineage and vector search. US20230061341A1 · Technion Research and Development Foundation Ltd. 2023-03-02 · Publicación de patente · Anterior al corte.

Aproxima el linaje de registros con vectores. La reivindicación 13 del A1 permite excluir registros posteriores o fuera del intervalo solicitado mediante la fecha de creación. Es un precedente directo de exclusión temporal. La fecha de creación y lo que realmente recibió el modelo necesitan evidencias distintas. Registro principal · US12346297B2 · 2025-07-01

M05 · Retrieving documents according to temporal similarity. US20230033211A1 · International Business Machines Corp. 2023-02-02 · Publicación de patente · Anterior al corte.

Modela relaciones temporales entre entidades para ordenar documentos por similitud temporal. Un resultado relevante puede haberse publicado después del corte. Registro principal · US12045270B2 · 2024-07-23

M06 · Large language model verification. US20250200392A1 · Amazon Technologies Inc. 2025-06-19 · Publicación de patente · Anterior al corte.

Traduce la respuesta a lógica de primer orden y la contrasta con fórmulas del dominio. El resultado depende de la traducción y los axiomas. Para auditar una consulta histórica se necesita además fechar el respaldo. Registro principal

M07 · Systems and methods for secure and fast machine learning inference in a trusted execution environment. US20220114249A1 · Huawei Technologies Co Ltd. 2022-04-14 · Publicación de patente · Anterior al corte.

Aprovisiona un modelo cifrado y protege su inferencia en un entorno de confianza. La atestación remota puede reforzar la evidencia del entorno utilizado. La fecha de cada fuente sigue necesitando una política verificable. Registro principal

M08 · Tracking machine learning data provenance via a blockchain. US20250200334A1 · Coinbase Inc. 2025-06-19 · Publicación de patente · Anterior al corte.

Registra en cadena la procedencia de los datos de desarrollo de un modelo. La unidad observada aquí es el conjunto de documentos de una única consulta. Registro principal

M09 · Semantic aware hallucination detection for large language models. US12387049B2 · Intuit Inc. 2025-08-12 · Publicación de patente · Anterior al corte.

Calcula la consistencia semántica entre preguntas equivalentes y respuestas de dos modelos. Ambos pueden coincidir en un dato sin respaldo público al corte. Intuit figura como solicitante. El A1 ya se había publicado en marzo de 2025. Registro principal · US20250077777A1 · 2025-03-06

M10 · Detecting hallucination in a language model. US20240419912A1 · Microsoft Technology Licensing LLC. 2024-12-19 · Publicación de patente · Anterior al corte.

Detecta material de ejemplos o conocimiento externo en una respuesta. Es pertinente cuando la propia pregunta contiene la palabra que dispara una alerta. Copiar un término no fecha una afirmación. Registro principal

M11 · Computer mechanism for analytics outcome transparency. US12099495B2 · International Business Machines Corp. 2024-09-24 · Publicación de patente · Anterior al corte.

Vincula los análisis acordados con resultados e historial de ejecución. El registro puede preservar una clasificación equivocada con la misma fidelidad que una correcta. Registro principal

M12 · Methods and apparatus to train a model using attestation data. US20230186156A1 · Intel Corp. 2023-06-15 · Publicación de patente · Anterior al corte.

Pondera datos de entrenamiento de dispositivos según atestación y telemetría. Para evaluar una salida histórica hacen falta además una afirmación y una referencia fechada que le corresponda. Registro principal

M13 · System and method for policy-constrained symbolic rendering of AI outputs. US12464027B2 · OneSource Solutions International Inc. 2025-11-04 · Publicación de patente · Anterior al corte.

Autoriza la representación simbólica de resultados en el subsistema gráfico. La reivindicación 1 vincula una cápsula de evidencia al símbolo y registra su justificación. Es un precedente para una decisión de política firmada. El conjunto documental suministrado se observa en una fase anterior. Registro principal

M14 · Blockchain ledger authentication. US20200195448A1 · Alibaba Group Holding Ltd. 2020-06-18 · Publicación de patente · Anterior al corte.

Autentica tramos de un registro con compromisos de Merkle y un servicio de tiempo. El sello puede fechar el compromiso, mientras que el suceso descrito necesita fuentes propias. Registro principal · US10778452B2 · 2020-09-15

M15 · Ledger verification method and apparatus, and device. US20200213094A1 · Alibaba Group Holding Ltd. 2020-07-02 · Publicación de patente · Anterior al corte.

Usa un certificado temporal para comprobar un tramo del registro. Esto respalda la cronología de registro. La interpretación de una afirmación de futuro necesita una evaluación distinta. Registro principal · US10790968B2 · 2020-09-29

M16 · Ledger verifiable-pruning system. EP4009579A2 · Bloom Technology Inc. 2022-06-08 · Publicación de patente · Anterior al corte.

Un árbol de Merkle sesgado permite podar registros conservando verificabilidad. La prueba reducida debe interpretarse respecto al compromiso de origen y las reglas de conservación. Registro principal · US11949801B2 · 2024-04-02

M17 · Methods and systems for facilitating collection of road user charges using a digital currency based on a distributed ledger technology. US20250378436A1. 2025-12-11 · Publicación de patente · Posterior al corte.

El documento trata del cobro por uso de carreteras mediante moneda digital. La serialización de transacciones es una coincidencia periférica. El A1 se publicó el 11 de diciembre de 2025 tras un depósito el 25 de agosto. Es una continuación en parte. Se enlazan las dos publicaciones anteriores de peajes. La búsqueda literal de canonical y Merkle en ambas no obtuvo coincidencias. Queda abierta la posible divulgación bajo otras expresiones. Registro principal · US20230196340A1 · US20220398870A1

M18 · Circuit and method for timestamp filtering with RLS filter. US12395417B2 · Marvell Asia Pte Ltd. 2025-08-19 · Publicación de patente · Anterior al corte.

Un filtro físico de mínimos cuadrados recursivos procesa marcas de reloj. Se ocupa de estimar una señal. El filtro documental decide disponibilidad histórica. La segunda referencia de Marvell es otra patente. Registro principal · US12267222B2 · 2025-04-01

M19 · A kind of event extraction method across media. CN106484767A · Institute of Information Engineering. 2017-03-08 · Publicación de patente · Anterior al corte.

Extrae elementos de sucesos en noticias y redes y los combina entre medios. Ayuda a vincular afirmaciones con sucesos. Falta determinar cuándo se conocían y qué afirma realmente el texto. Registro principal · CN106484767B · 2019-06-21

Identidad, procedencia de activos y computación verificable

M20 · Key Event Receipt Infrastructure (KERI): Design and Build. KERI white paper v1.6 · Samuel M. Smith. 2019-07-03 · Fecha del documento técnico · Anterior al corte.

KERI conserva eventos firmados de claves y recibos para reconstruir el control de un identificador. El historial de una clave por sí solo no establece la veracidad de la respuesta firmada. Registro principal

M21 · C2PA technical specifications, version 1.0. C2PA 1.0 · Coalition for Content Provenance and Authenticity. 2022-01-26 · Anuncio de la organización · Anterior al corte.

Los manifiestos C2PA vinculan declaraciones y procedencia a contenidos digitales. El valor de una declaración firmada depende del firmante, de los campos incluidos y del modelo de confianza. Se usa la versión 1.0 para fijar fecha. Registro principal · C2PA 1.0

M22 · Distributed Ledger for Provenance Tracking of Artificial Intelligence Assets. arXiv:2002.11000 · Philipp Lüthi, Thibault Gagnaux, Marcel Gygli. 2020-02-25 · Primer envío en arXiv · Anterior al corte.

Emplea un registro distribuido para seguir conjuntos de datos y modelos. Aporta linaje de desarrollo. Una auditoría temporal examina una afirmación fechada de una ejecución concreta. Registro principal

M23 · EZKL, release v12.0.0. EZKL v12.0.0 · EZKL / zkonduit. 2024-07-24 · Versión publicada del software · Anterior al corte.

EZKL permite generar y verificar pruebas de un grafo de cálculo. Se puede acreditar el cálculo declarado. Para acreditar la admisibilidad de una fuente hay que incorporarla al enunciado o a sus entradas de confianza. Registro principal

M24 · RISC Zero zkVM, release v1.0.0. RISC Zero zkVM v1.0.0 · RISC Zero. 2024-06-04 · Versión publicada del software · Anterior al corte.

RISC Zero ofrece una máquina virtual de conocimiento cero. Un recibo vincula el programa con sus salidas públicas. La interpretación temporal depende de las comprobaciones programadas y de los datos recibidos. Registro principal

M25 · zkLLM: Zero Knowledge Proofs for Large Language Models. arXiv:2404.16109 · Haochen Sun, Jason Li, Hongyang Zhang. 2024-04-24 · Primer envío en arXiv · Anterior al corte.

zkLLM desarrolla pruebas de conocimiento cero para la inferencia de transformadores. Verificar una operación sobre entradas concretas deja abierta la interpretación histórica de la respuesta. Registro principal

M26 · Scaling up Trustless DNN Inference with Zero-Knowledge Proofs. arXiv:2210.08674 · Daniel Kang, Tatsunori Hashimoto, Ion Stoica, Yi Sun. 2022-10-17 · Primer envío en arXiv · Anterior al corte.

Escala las pruebas de ejecución de redes neuronales. La corrección computacional puede ser objeto de una prueba. Esa garantía presupone una operación especificada y no certifica por sí sola la fecha de un documento. Registro principal

M27 · Verifiable evaluations of machine learning models using zkSNARKs. arXiv:2402.02675 · Tobin South, Alexander Camuto, Shrey Jain, Shayla Nguyen, Robert Mahari, Christian Paquin, Jason Morton, Alex 'Sandy' Pentland. 2024-02-05 · Primer envío en arXiv · Anterior al corte.

Usa zkSNARKs para comprobar evaluaciones de modelos con material confidencial. Una puntuación verificable aún puede depender de etiquetas que confunden una predicción con una afirmación. Registro principal

M28 · Attestable Audits: Verifiable AI Safety Benchmarks Using Trusted Execution Environments. arXiv:2506.23706 · Christoph Schnabl, Daniel Hugenroth, Bill Marino, Alastair R. Beresford. 2025-06-30 · Primer envío en arXiv · Anterior al corte.

Attestable Audits usa hardware de confianza para pruebas de seguridad verificables. Atestiguar qué evaluación se ejecutó deja abierta la discusión sobre la validez de sus etiquetas temporales. Registro principal

M29 · Trustless Audits without Revealing Data or Models. arXiv:2404.04500 · Suppakit Waiwitlikhit, Ion Stoica, Yi Sun, Tatsunori Hashimoto, Daniel Kang. 2024-04-06 · Primer envío en arXiv · Anterior al corte.

zkAudit estudia cómo auditar datos de entrenamiento y modelos reservados. Es pertinente para evaluaciones reproducibles con entradas restringidas. La interpretación de afirmaciones fechadas requiere reglas adicionales. Registro principal

M30 · Prove AI governance product announcement (Casper Labs). Prove AI announcement · Casper Labs. 2024-05-22 · Anuncio de la organización · Anterior al corte.

Casper Labs anunció un producto de gobernanza y procedencia de datos. El anuncio documenta fecha y funcionalidad prevista. No valida por separado el rendimiento ni la integridad de cada consulta. Registro principal

M31 · AICert v1.0: Open-Source AI Traceability Tool for Verifiable Training. AICert v1.0 · Raphaël Millet / Mithril Security. 2024-09-30 · Anuncio de la organización · Anterior al corte.

AICert v1.0 describe entrenamiento verificable con computación confidencial y evidencia firmada. La procedencia del modelo y el contenido entregado en cada consulta necesitan observaciones distintas. Registro principal

Recuperación temporal y grafos de sucesos

M32 · It's About Time: Incorporating Temporality in Retrieval Augmented Language Models. arXiv:2401.13222 · Anoushka Gade, Jorjeta Jetcheva. 2024-01-24 · Primer envío en arXiv · Anterior al corte.

TempRALM incorpora la pertinencia temporal al recuperar documentos para una pregunta. Sirve de referencia de recuperación. La evaluación forense de cada afirmación fechada es otra tarea. Registro principal

M33 · TimeR4: Time-aware Retrieval-Augmented Large Language Models for Temporal Knowledge Graph Question Answering. 10.18653/v1/2024.emnlp-main.394 · Xinying Qian, Ying Zhang, Yu Zhao, Baohang Zhou, Xuhui Sui, Li Zhang, Kehui Song. 2024-11 · Ficha de publicación · Anterior al corte.

TimeR4 combina recuperación temporal y modelos de lenguaje para preguntas sobre grafos temporales. Acertar una respuesta mide desempeño y por sí solo no identifica el canal de conocimiento que produjo el acierto. Registro principal

M34 · Mitigating Temporal Misalignment by Discarding Outdated Facts. arXiv:2305.14824 · Michael J.Q. Zhang, Eunsol Choi. 2023-05-24 · Primer envío en arXiv · Anterior al corte.

Usa la duración estimada de hechos para descartar información volátil o anticuada y mejorar la calibración. La volatilidad predicha estima la vida útil de un dato, mientras que una instantánea registra una versión concreta. Registro principal · EMNLP 2023

M35 · Causal Event Graph-Guided Language-based Spatiotemporal Question Answering. 10.1609/aaaiss.v3i1.31204 · Kaushik Roy, Alessandro Oltramari, Yuxin Zi, Chathurangi Shyalika, Vignesh Narayanan, Amit P. Sheth. 2024-05-20 · Ficha de publicación · Anterior al corte.

Modela relaciones causales entre sucesos para contestar preguntas espaciotemporales. Una arista entre sucesos expresa una relación de razonamiento. La disponibilidad histórica de cada dato es otra propiedad. Registro principal

M36 · Two-stage Generative Question Answering on Temporal Knowledge Graph Using Large Language Models. arXiv:2402.16568 · Yifu Gao, Linbo Qiao, Zhigang Kan, Zhihua Wen, Yongquan He, Dongsheng Li. 2024-02-26 · Primer envío en arXiv · Anterior al corte.

GenTKGQA recupera un subgrafo y después genera una respuesta temporal. Es un precedente de recuperación y generación orientadas al tiempo. Esa arquitectura por sí sola no emite un recibo del conjunto exacto de entradas. Registro principal · Findings of ACL 2024

Registros de agentes, recuperación en grafos y memoria temporal

M37 · ALETHEIA Protocol: a decision log for AI agents. Esrbwt1/Aletheia-Protocol · Esrbwt1. 2025-08-26 · Metadatos de commit · Solo historial de commits.

El proyecto describe un historial encadenado por hash de decisiones de agentes. Los tres commits registrados son del 26 de agosto de 2025. Acreditan el historial de código. Falta una captura independiente de cuándo se hizo público inicialmente. Registro principal · 2025-08-26 · 5bac41e0af7

M38 · Temporal GraphRAG hard filter (unresolved lead). Unresolved Rust module. Fuente sin confirmar · Identidad y fecha sin resolver.

Aparecía la pista de un módulo Rust llamado temporal_retrieval. El candidato documenta intervalos y datos de grafo con validez temporal. No se confirmó la identidad, fecha ni existencia en esa versión del supuesto filtro excluyente. graphrag-core 0.1.1 (?)

M39 · CausalRAG: Integrating Causal Graphs into Retrieval-Augmented Generation. arXiv:2503.19878 · Nengbo Wang, Xiaotian Han, Jagdip Singh, Jing Ma, Vipin Chaudhary. 2025-03-25 · Primer envío en arXiv · Anterior al corte.

CausalRAG usa la estructura causal de un grafo para mejorar la recuperación. Los vínculos del grafo no atribuyen el origen de una frase generada. La v1 de arXiv y la publicación de ACL Findings son anteriores al corte. Registro principal · Findings of ACL 2025

M40 · CDF-RAG: Causal Dynamic Feedback for Adaptive Retrieval-Augmented Generation. arXiv:2504.12560 · Elahe Khatibi, Ziyu Wang, Amir M. Rahmani. 2025-04-17 · Primer envío en arXiv · Anterior al corte.

CDF-RAG afina preguntas y contrasta respuestas con rutas causales recuperadas. Une búsqueda y validación. La admisibilidad temporal y el recibo verificable del contexto suministrado requieren controles propios. Registro principal

M41 · Respecting Temporal-Causal Consistency: Entity-Event Knowledge Graphs for Retrieval-Augmented Generation. arXiv:2506.05939 · Ze Yu Zhang, Zitao Li, Yaliang Li, Bolin Ding, Bryan Kian Hsiang Low. 2025-06-06 · Primer envío en arXiv · Anterior al corte.

E²RAG relaciona entidades y sucesos para mejorar la consistencia temporal y causal. ChronoQA evalúa preguntas sensibles al tiempo. La coherencia del razonamiento y la autenticación del conjunto de entradas miden propiedades distintas. Registro principal

M42 · Beyond Single Pass, Looping Through Time: KG-IRAG with Iterative Knowledge Retrieval. arXiv:2503.14234 · Ruiyi Yang, Hao Xue, Imran Razzak, Hakim Hacid, Flora D. Salim. 2025-03-18 · Primer envío en arXiv · Anterior al corte.

KG-IRAG repite la recuperación sobre un grafo temporal. Puede mejorar las respuestas históricas aunque no se disponga de un historial fechado por terceros de los registros admitidos para cada salida. Registro principal

M43 · Zep: A Temporal Knowledge Graph Architecture for Agent Memory. arXiv:2501.13956 · Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, Daniel Chalef. 2025-01-20 · Primer envío en arXiv · Anterior al corte.

Zep mantiene relaciones temporales en la memoria de un agente. Diferenciar el cambio real del momento de registro ayuda a elegir qué fecha usar. El grafo sigue dependiendo de la procedencia de esas fechas. Registro principal

Alucinaciones y señales de respaldo contextual

M44 · Unsupervised Real-Time Hallucination Detection based on the Internal States of Large Language Models. arXiv:2403.06448 · Weihang Su, Changyue Wang, Qingyao Ai, Yiran HU, Zhijing Wu, Yujia Zhou, Yiqun Liu. 2024-03-11 · Primer envío en arXiv · Anterior al corte.

MIND extrae indicios de alucinación de estados internos sin etiquetas supervisadas. Puede complementar las referencias históricas cuando se dispone de acceso a las activaciones necesarias. Registro principal

M45 · Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using Only Attention Maps. arXiv:2407.07071 · Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, James Glass. 2024-07-09 · Primer envío en arXiv · Anterior al corte.

Lookback Lens usa la atención al contexto y a tokens ya generados para detectar alucinaciones contextuales. Leer esos pesos no verifica por separado las fechas de publicación del material recuperado. Registro principal

M46 · REFIND at SemEval-2025 Task 3: Retrieval-Augmented Factuality Hallucination Detection in Large Language Models. arXiv:2502.13622 · DongGeon Lee, Hwanjo Yu. 2025-02-19 · Primer envío en arXiv · Anterior al corte.

REFIND usa evidencia recuperada y sensibilidad al contexto para localizar fragmentos sin respaldo. La evaluación histórica exige distinguir además hechos posteriores de previsiones legítimas. Registro principal

M47 · Prompt-Guided Internal States for Hallucination Detection of Large Language Models. 10.18653/v1/2025.acl-long.1058 · Fujie Zhang, Peiqi Yu, Biao Yi, Baolei Zhang, Tong Li, Zheli Liu. 2025-07 · Ficha de publicación · Anterior al corte.

Preguntas diseñadas para revelar estados internos alimentan un clasificador de alucinaciones. Depende de acceso al modelo y datos de calibración. El recibo de una ejecución y la cronología de sus fuentes responden a otras preguntas. Registro principal

Registros de transparencia, canonicalización y repetibilidad

M48 · Certificate Transparency. RFC 6962 · B. Laurie, A. Langley, E. Kasper. 2013-06 · Publicación del RFC · Anterior al corte.

Certificate Transparency especifica un registro de Merkle y pruebas de consistencia e inclusión. Permite entender qué compromete una raíz. La integridad del registro no establece la veracidad de sus entradas. El RFC 9162 lo reemplaza. Registro principal · RFC 9162

M49 · JSON Canonicalization Scheme (JCS). RFC 8785 · A. Rundgren, B. Jordan, S. Erdtman. 2020-06 · Publicación del RFC · Anterior al corte.

JCS define serialización canónica de JSON para reproducir los bytes firmados. Cumplir el estándar exige aplicar sus reglas de cadenas y números. La mera ordenación de claves no basta. Registro principal

M50 · RepDL: Bit-level Reproducible Deep Learning Training and Inference. RepDL · Peichen Xie, Xian Zhang, Shuo Chen. 2025-10 · Ficha de publicación · Anterior al corte.

RepDL busca igualdad a nivel de bits en entrenamiento e inferencia bajo configuraciones admitidas. Repetir operaciones numéricas y reconstruir qué evidencia histórica estaba disponible son objetivos complementarios. Registro principal · arXiv:2510.09180 · 2025-10-10

Trabajos posteriores y RAG forense en el ámbito jurídico

M51 · AuditableLLM: A Hash-Chain-Backed, Compliance-Aware Auditable Framework for Large Language Models. 10.3390/electronics15010056 · Delong Li, Guangsheng Yu, Xu Wang, Bin Liang. 2025-12-23 · Ficha de publicación · Posterior al corte.

AuditableLLM separa la actualización del modelo de su capa de auditoría y encadena por hash los registros sucesivos de actualización. La unidad observada difiere de una pregunta y su respuesta. La revista publicó la versión definitiva el 23 de diciembre de 2025. No se acreditó aquí una publicación pública anterior. Registro principal · Version history

M52 · Autonomous Action Runtime Management (AARM): A System Specification for Securing AI-Driven Actions at Runtime. arXiv:2602.09433 · Herman Errico. 2026-02-10 · Primer envío en arXiv · Posterior al corte.

AARM especifica cómo clasificar y autorizar acciones de agentes durante la ejecución. Es un desarrollo posterior de gobernanza. La política de acciones no proporciona las etiquetas históricas de una prueba factual. Registro principal

M53 · RAGShield: Detecting Numerical Claim Manipulation in Government RAG Systems. arXiv:2604.00387 · KrishnaSaiReddy Patil. 2026-04-01 · Primer envío en arXiv · Posterior al corte.

RAGShield contrasta afirmaciones numéricas con un registro multifuente y detecta cambios fuera de calendario. Ilustra después del corte la utilidad de definir con precisión el hecho y el dominio. Su ámbito son datos públicos numéricos. Registro principal

M54 · Gobernanza y trazabilidad "a prueba de AI Act" para casos de uso legales: un marco técnico-jurídico, métricas forenses y evidencias auditables. arXiv:2510.12830 · Alex Dantart. 2025-10-12 · Primer envío en arXiv · Anterior al corte.

La v1 de Dantart ya selecciona material jurídico por vigencia y firma un registro WORM con pregunta, contexto temporal, huellas de documentos, respuesta y decisión de política. Es la comparación más próxima del catálogo inicial. La v2 llegó en diciembre. Registro principal · v1 §§3.2, 5.3 · v2 · 2025-12-23

Cobertura de la revisión y procedencia de las referencias

Qué se contó

La lista consolidada de mayo tenía **54 posiciones**. Dos evaluaciones conservadas además en formato de texto y en documento de oficina aportaban **36 y 46 citas bibliográficas**. Se contrastaron ambas versiones de cada lista. Sus 82 posiciones se corresponden con **72 registros** después de reunir versiones y enlaces alternativos. Otros **44 registros** proceden del texto de los estudios, del material de búsqueda y seguimiento, de las notas técnicas y jurídicas, o de fuentes primarias localizadas para identificar afirmaciones que allí figuraban. El anexo bibliography.csv contiene así **116 registros**. B052 y M39 corresponden al mismo trabajo. Los totales no se pueden sumar como si fueran estudios o patentes distintos.

La revisión de mayo, las dos listas de fuentes, sus referencias suplementarias y la búsqueda posterior emplean distintas unidades de recuento. Las 54 posiciones originales figuran en legacy.csv. La columna `historical_locations` conserva los números de referencia de las dos evaluaciones, también para las citas repetidas. Se distinguen los trabajos técnicos, las patentes, las páginas de proyectos en desarrollo, las

actuaciones de organismos públicos, los comentarios y los nombres sin resolver por su tipo y grado de verificación. Para contar estudios técnicos comparables es necesario filtrar los registros por tipo y alcance.

Denominadores por documento

Los códigos identifican grupos del material de trabajo para poder auditar las cifras sin depender de documentos privados. Para archivos sin bibliografía enumerada se indica qué se podía contar exactamente.

Código	Material	Denominador y tratamiento
S	Síntesis consolidada de los trabajos de búsqueda, mayo de 2026	54/54 filas numeradas, M01–M54. La tabla adicional de §3.4 tiene 11 filas y la de §8 tiene 14. Son 14 nombres distintos entre ambas, porque las 11 entradas de §3.4 reaparecen en §8. Se concilian más abajo.
F	Auditoría forense de IA, evaluación técnica y comercial	36/36 posiciones bibliográficas, F01–F36. El documento de oficina contiene 36 destinos de enlace externos y la versión de texto enumera 36 entradas. Representan 31 registros. Se identificaron además fuentes para términos y actuaciones citados en el cuerpo.
P	Evaluar la patentabilidad de la auditoría de IA, evaluación de entorno y mercado	46/46 posiciones bibliográficas, P01–P46. El documento de oficina contiene 46 destinos de enlace externos y el texto conserva 46 entradas. Representan 41 registros. Las páginas de tarifas se retienen por trazabilidad histórica.
Q0	Diseño de la búsqueda	Cuatro frentes y diez líneas de referencias ya conocidas, una de las cuales enumera dos patentes. Aporta consultas y nombres para comprobar. No constituye un estudio adicional de cada resultado.
Q-C	Análisis técnico multieje	18 subsecciones temáticas antes de la tabla de síntesis, incluidas ocho sobre patentes. Nueve apariciones de URL de texto con siete destinos distintos, varias corresponden a páginas institucionales genéricas. Las fichas de patente, las publicaciones académicas y las normas identificables se remiten a M o B. La mención de Giza y Aleo se conserva en B095–B096.
Q-D	Primer informe complementario de búsqueda	Ocho encabezados de hallazgos, cuatro de ellos sobre patentes. Las otras entradas abarcan EZKL/RISC Zero/ZkAudit, Prove AI, TempRALM y ALETHEIA. Sus identificadores o las limitaciones de procedencia se recogen en M23–M24, M29–M30, M32 y M37.
Q-G	Segundo informe complementario de búsqueda	Diez encabezados temáticos. Los nombres que quedaban fuera de las 54 filas incluyen Jolt, vCNN, MECM, KEA y la mención de Qiu. B003, B087, B089, B090 y B094 conservan el resultado.
Q-P	Tercer informe complementario de búsqueda	Seis encabezados de hallazgos individuales, cuatro sobre patentes individuales o familias. La copia de trabajo menciona productos y fuentes mediante texto, pero no contiene un inventario de enlaces reproducible. M13–M16, M21–M22 y las respectivas familias recogen los encabezados.
Q-T	Informe transversal de convergencia	Dos grupos nombrados de coincidencias de componentes y dos hipótesis de combinación. Se conservan las referencias a KEA y MECM. Las coincidencias de vocabulario y las búsquedas sin resultado no sustituyen una comparación de reivindicaciones.

Código	Material	Denominador y tratamiento
T	Seguimiento técnico, sección B	Cuatro apartados. B1 remite a IBM M01, B2 a SEAL, B3 a NAL y B4 añade cuatro trabajos sobre entornos de ejecución, otro sobre integridad de contexto, una ficha comercial y cinco identificadores de patentes. El texto hablaba de diez patentes, pero solo enumeraba cinco. B082–B086 contienen esas cinco, sin inferir la dirección del vínculo de cita ni identificar las otras cinco.
D1–D4	Seguimiento, especificación técnica, borrador de reivindicaciones dependientes y complemento de ingeniería	Cuatro documentos, incluido T. Los dos textos de ingeniería cubren los mismos cuatro temas (marcas de tiempo, límites de ejecución, truncamiento y versión del modelo). RFC 3161 se identifica en B097. Los detalles de diseño son propuestas del proyecto.
L1	Discusión de elegibilidad de materia patentable en EE. UU.	Siete entradas en la tabla de casos, más la decisión Alice en el cuerpo. B102–B109 identifican las ocho. Sirven para documentar el contexto jurídico de la revisión.
L2	Nota de continuidad técnica	Seis áreas de búsqueda y tres nombres en el apartado de convergencia. B047, B050 y B052 corrigen la identidad, fecha o alcance.
E	Nota de metodología de evaluación temporal	Cinco apartados de desarrollo de la metodología. Sin bibliografía definida. Un ejemplo de aplicación citado allí no es una validación independiente.
Contexto de marcas	Conjunto de estudios conservado con F y P	33 archivos en la carpeta estudiada, incluidos los cuatro archivos correspondientes a F y P. Los otros 29 textos se cribaron buscando identificadores reconocibles de publicaciones académicas y patentes, DOI, RFC y enlaces de investigación. El cribado no devolvió coincidencias. Los nombres sin identificador pudieron escapar al filtro. Los casos empresariales no se contabilizan como antecedentes técnicos.

Conciliación de las dos tablas secundarias del informe consolidado

Entrada histórica	Registro(s)	Tratamiento cronológico y técnico
AuditableLLM	M51	Publicación posterior, sobre trazas de interacciones.
Agentic Witnessing y Aegon	B020 y B005	Trabajos de abril de 2026. La identidad del testigo, el acceso al contenido o un registro de agente responden a preguntas distintas de la fecha en que estuvo disponible un hecho.
HugRAG	B050	v1 de febrero de 2026. El propio registro se tituló CausalRAG2 posteriormente. B052/M39 es otro trabajo, de marzo de 2025.
Semantic Density Effect	B047	Abril de 2026. Estudia información por token en las instrucciones al modelo. B003, de 2024, estudia incertidumbre de respuestas en un espacio semántico.
ZKPROV y VeriLLM	B073 y B074	Los depósitos v1 son del 26 de junio y del 29 de septiembre de 2025. La versión 2 de ZKPROV sí es posterior a la fecha de comparación. El rótulo de la antigua tabla inducía a error en ambos casos.
MathLedger	B075	Depósito v1 el 22 de diciembre de 2025. El prefijo 2601 de arXiv no da la fecha exacta de depósito.
GRAIL Protocol	B093	Etiqueta sin identificación inequívoca. No fundamenta ninguna comparación.
AARM y RAGShield	M52 y M53	Ambos de 2026. RAGShield trata la manipulación de afirmaciones numéricas, una cuestión más concreta que la descrita en la síntesis histórica.
Zep, KG-IRAG y CDF-RAG	M43, M42 y M40	Las tres primeras versiones son de enero, marzo y abril de 2025, respectivamente. Aparecían en §8 pese a ser anteriores al 4 de diciembre.

Qué se verificó

El CSV registra identificadores de publicación, páginas primarias o URL consignadas, el estado de verificación, la fecha de la edición identificada y un límite interpretativo por entrada. Los títulos de arXiv normalmente son los actuales. El campo de fecha sí corresponde al primer depósito cuando el historial permitió establecerlo. El texto de una revisión posterior necesita su propia fecha. La fecha de una patente es la fecha pública de la publicación concreta indicada. El análisis de familia y de fechas efectivas exige revisar otros documentos. El campo vacío conserva la incertidumbre.

De los 116 registros, 46 incluyen una fecha anterior al 4 de diciembre de 2025, 12 una posterior y 58 carecen de datación histórica adecuada. Estas cifras reúnen materiales heterogéneos. Incluyen la fecha atribuida por el propio editor para NAL, las fechas de escritos judiciales, documentos administrativos y obras técnicas. Las dos denominaciones sin una fuente inequívoca son GRAIL y MECM. Las páginas de catálogos y herramientas pueden cambiar con el tiempo.

La verificación distingue 22 registros arXiv cotejados con historial y ocho publicaciones de patente con título, número y fecha pública contrastados. Se revisaron también identificadores y catálogos específicos de otras publicaciones, normas y decisiones. En 39 entradas contextuales se pudo consultar una página sin establecer qué versión existía en mayo. Quince conservan el destino registrado sin confirmación independiente del contenido o de la versión. La columna de estado conserva otras situaciones, como una decisión identificada o una página comercial sin edición histórica. No se leyeron íntegramente todos los textos de la bibliografía.

Agrupamos por DOI, identificador de preprint, número de publicación de patente o URL documental. La sede con y sin www y la ficha y el PDF de la misma publicación pueden figurar juntos. Para P41 y P46 se reunieron dos nombres de archivo del mismo organismo bajo el rótulo de tarifas común, sin verificar igualdad binaria o revisión histórica. Los formatos y las URLs originales adicionales se conservan cuando aportan trazabilidad.

Qué aporta al análisis del método

- **Recuperación temporal.** TA-RAG y STAR-RAG (B002 y B001) estudian preguntas diacrónicas y selección de evidencia mediante grafos. Las medidas de recuperación o exactitud de respuesta no verifican por sí solas qué bytes recibió una petición concreta. TV-RAG (B004) incorpora organización temporal y pesos de entropía para vídeo. El orden de fotogramas y la fecha pública de un documento son variables distintas.
- **Incertidumbre.** Qiu y Miikkulainen (B003), Farquhar y coautores (B088), el trabajo de incertidumbre geométrica (B015) y KEA Explain (B087) ofrecen instrumentos para estudiar señales de alucinación. Las respuestas pueden coincidir y ser erróneas. Un escalar no permite identificar por sí mismo la causa de una respuesta ni el momento en que se conoció un hecho. Youtu-VL (B049) tampoco sustenta la afirmación independiente sobre JEPA con la que aparecía asociado.
- **Ejecución.** Jolt y vCNN tratan la verificación de cómputo. SecureInfer y TZ-LLM incorporan entornos de ejecución protegidos anteriores al corte. AgenTEE, FlexServe y Protecting Context and Prompts son comparadores posteriores. SEAL vincula marcas de agua y adaptación de modelos. Conviene especificar qué proceso consume los bytes comprometidos, qué versión se identifica y quién acredita el tiempo. Las notas de ingeniería también nombran la integridad de memoria física y mecanismos de refresco de memoria (TRR). Son hipótesis de entorno, sin prueba experimental en este estudio.
- **Compromisos y fechas.** Las ocho publicaciones de patente del anexo ilustran piezas como registros, validación de múltiples partes o comprobación de pertenencia mediante Merkle. Pertenecer al conjunto comprometido, aparecer en el historial de ingestión y ser de conocimiento público antes de un corte son tres predicados distintos. RFC 3161 vincula un resumen con una marca de tiempo bajo la confianza depositada en la autoridad. DER (B110) fija reglas de codificación. HNSW (B111) contextualiza la recuperación aproximada y por qué una búsqueda puede omitir evidencia elegible.
- **Mercado y supervisión.** F y P hablan de gestión de cumplimiento y auditoría (Vanta, Secureframe, SafeBase, Bonsai Security, Finspectors, MindBridge y Caseware), seguridad de plataformas (AccuKnox, SentinelOne, Traceable y Levo) y visibilidad de marca (Sight AI, Evertune, Indexly, Peec AI, Waikay, Semrush, Ahrefs, BrightEdge y Writesonic). En general el respaldo bibliográfico son listas comerciales. No se evaluaron esos productos con el protocolo. La mención del proveedor Modulus Labs en los materiales de búsqueda carecía de edición técnica identificada. B112–B115 permiten distinguir actuaciones de la SEC, el DOJ y la FTC. B116 identifica una demanda de derechos de autor. Alegaciones, acuerdos, directrices y normas tienen peso jurídico distinto. El apartado europeo del artículo remite al texto legal aplicable.
- **Qué queda abierto.** La revisión conserva las omisiones identificables y los casos ambiguos. No permite inferir que todos los trabajos públicos están incluidos, que nadie publicó antes una combinación, que implementar el método garantiza conformidad con el Reglamento de IA, ni que se verificó la verdad o el origen causal de una respuesta.

Cómo reconstruir la revisión

Ordenando las marcas F01–F36 y P01–P46 se recuperan las dos bibliografías. Sus 72 registros son B001–B072. Las entradas B073–B116 añaden comparaciones técnicas, información contextual, identificadores ambiguos y fuentes primarias de afirmaciones del cuerpo de los informes. Solo B052 coincide explícitamente con la identidad de un registro M numerado (M39). Otros registros pueden abordar un mismo mecanismo. Los recuentos corresponden a esta edición del anexo y las fechas se compararon con el 4 de diciembre de 2025. La clasificación temporal es bibliográfica. La calificación jurídica exige otro análisis.

Las correcciones cronológicas afectan a la interpretación de los encabezados originales. ZKPROV y VeriLLM tenían preprints anteriores a la fecha de comparación. El primer depósito de MathLedger es posterior. HugRAG comparte registro con su versión retitulada CausalRAG2, mientras que CausalRAG (M39) es otra obra. El cambio de versión también afecta al título y a la interpretación de Deterministic Legal Agents (B030). El corpus abarca la relevancia temporal de una búsqueda, la incertidumbre del modelo, la integridad del cálculo, la pertenencia a un conjunto de datos y la disponibilidad pública independiente. Para cada referencia, la ficha permite identificar cuál de esas preguntas aborda y qué prueba adicional requerirían las demás.

Catálogo de las 116 referencias reconciliadas y complementarias

Los identificadores B001–B072 conservan las 82 posiciones de las dos bibliografías originales. B073–B116 recuperan referencias adicionales o aclaran sus fuentes. La nota de cada ficha explica su función. Se mantiene el título en su lengua original. La procedencia conserva los ordinales, incluso al repetirse una fuente.

B001–B031

B001 · Right Answer at the Right Time - Temporal Retrieval-Augmented Generation via Graph Summarization. arXiv:2510.16715 · investigación · 2025-10-19.

El método resume un grafo para recuperar evidencia pertinente a lo largo del tiempo. Evalúa la selección de documentos. Queda fuera de su objetivo comprobar qué secuencia exacta de bytes se envió al modelo. Procedencia F01, F06, Q0. Identificador primario e historial de versiones comprobados. Ficha Otra fuente 1

B002 · Reading Between the Timelines: RAG for Answering Diachronic Questions. arXiv:2507.22917 · investigación · 2025-07-21.

Los metadatos temporales y la cobertura de intervalos ayudan a responder preguntas sobre cambios a lo largo del tiempo. La pertinencia respecto a una fecha no garantiza la exclusión de todos los documentos posteriores. Procedencia F02, Q0. Identificador primario e historial de versiones comprobados. Ficha Otra fuente 1

B003 · Semantic Density: Uncertainty Quantification for Large Language Models through Confidence Measurement in Semantic Space. arXiv:2405.13845 · investigación · 2024-05-22.

Estima la incertidumbre a partir de la densidad semántica de las respuestas. La confianza de un modelo no permite determinar cuándo se conoció un hecho ni de dónde proviene. Procedencia F03, F16, Q-G/Q0/S6. Identificador primario e historial de versiones comprobados. Ficha Otra fuente 1

B004 · TV-RAG: A Temporal-aware and Semantic Entropy-Weighted Framework for Long Video Retrieval and Understanding. arXiv:2512.23483 · investigación · 2025-12-29.

Combina la estructura temporal del vídeo con ponderación por entropía para recuperar segmentos de vídeos largos. La posición de un segmento dentro de la grabación y la fecha de publicación de una fuente son variables diferentes. Procedencia F04, F05, Q0. Identificador primario e historial de versiones comprobados. Ficha Otra fuente 1

B005 · Aegon: Auditable AI Content Access with Ledger-Bound Tokens and Hardware-Attested Mobile Receipts. arXiv:2604.06693 · investigación · 2026-04-08.

Aegon registra accesos mediante tokens vinculados a un registro y recibos atestados por el dispositivo móvil. Acreditar un acceso concreto deja abierta la cuestión de cuándo estuvo disponible públicamente el hecho descrito. Procedencia F07, F24, S3.4/S8. Identificador primario e historial de versiones comprobados. Ficha Otra fuente 1

B006 · Antony Rowstron's research works | Microsoft and other places - ResearchGate. pista de búsqueda · sin fecha acreditada.

Perfil bibliográfico de Antony Rowstron, conservado como pista para localizar publicaciones. La página actual del perfil no acredita la primera fecha de publicación de una obra. Procedencia F08. Pista registrada. Sin confirmación independiente del contenido o de la versión. Ficha

B007 · AI Washing | The New Frontier of Corporate Scrutiny. comentario sobre actuaciones · sin fecha acreditada.

Comentario sobre el escrutinio de afirmaciones comerciales relacionadas con la IA. Para establecer que hubo un procedimiento concreto se necesita la fuente del organismo competente. Procedencia F09. Página consultada. Sin verificación histórica. Ficha

B008 · One Year In, FTC's “Operation AI Comply” Continues Under New Administration, Signaling Enduring Enforcement Focus - Benesch Law. comentario sobre actuaciones · sin fecha acreditada.

Análisis jurídico de la operación AI Comply. Aporta contexto sobre la actuación de la FTC. La existencia y el alcance de cada procedimiento deben cotejarse en sus documentos oficiales. Procedencia F10. Página consultada. Sin verificación histórica. Ficha

B009 · 2025 Cybersecurity and AI Year in Review | Insights - Holland & Knight. comentario sobre actuaciones · sin fecha acreditada.

Balance de actuaciones en ciberseguridad e IA. Es una fuente de contexto regulatorio. La constatación de un procedimiento concreto requiere el expediente del regulador. Procedencia F11. Página consultada. Sin verificación histórica. Ficha

B010 · FINOS AI Governance Framework:. proyecto sectorial · sin fecha acreditada.

Marco de gobernanza de FINOS. En esta revisión no se evaluó un detector temporal del proyecto con versión fechada y replicación independiente. Procedencia F12. Página consultada. Sin verificación histórica. Ficha

B011 · The Biteris Protocol: The Global Backbone of Trusted Artificial Intelligence - Medium. proyecto sectorial · sin fecha acreditada.

Presentación del protocolo Biteris como proyecto de confianza y gobernanza. No se examinó un detector temporal de versión fechada y replicado de forma independiente. Procedencia F13. Pista registrada. Sin confirmación independiente del contenido o de la versión. Ficha

B012 · Structure-Aware Temporal Graph RAG - Emergent Mind. pista de búsqueda · sin fecha acreditada.

Ficha secundaria que permite localizar trabajos sobre RAG y grafos temporales. Su fecha actual carece de valor probatorio sobre la primera publicación del mecanismo. Procedencia F14. Página consultada. Sin verificación histórica. Ficha

B013 · Uncertainty quantification in LLMs. pista de búsqueda · sin fecha acreditada.

Recurso divulgativo para localizar métodos de cuantificación de incertidumbre. La página consultada no permite determinar cuándo se publicó por primera vez cada método. Procedencia F15. Pista registrada. Sin confirmación independiente del contenido o de la versión. Ficha

B014 · Quantifying uncertainty in LLMs with Semantic Density | by Cognizant AI Lab | Medium. pista de búsqueda · sin fecha acreditada.

Artículo divulgativo sobre densidad semántica. Orienta hacia el trabajo de investigación. Su versión actual no acredita la fecha de primera divulgación del método. Procedencia F17. Pista registrada. Sin confirmación independiente del contenido o de la versión. Ficha

B015 · Geometric Uncertainty for Detecting and Correcting Hallucinations in LLMs. arXiv:2509.13813 · investigación · 2025-09-17.

Estima la incertidumbre geométrica a partir de representaciones de las salidas. Puede informar una puntuación de alucinación, pero no permite fechar retrospectivamente la información. Procedencia F18. Fuente primaria pública identificada. Ficha Otra fuente 1

B016 · EdinburghNLP/awesome-hallucination-detection - GitHub. pista de búsqueda · sin fecha acreditada.

Lista de recursos sobre detección de alucinaciones. Sirve para localizar estudios. La página actual del repositorio no acredita su primera fecha de publicación. Procedencia F19. Página consultada. Sin verificación histórica. Ficha

B017 · UQLM: Uncertainty Quantification for Language Models, is a Python package for UQ-based LLM hallucination detection - GitHub. software · sin fecha acreditada.

UQLM reúne software y métodos de referencia para cuantificar la incertidumbre. La revisión no pudo fijar una versión histórica concreta del paquete. Procedencia F20. Página consultada. Sin verificación histórica. Ficha

B018 · Architectures, systems and methods for program defined state system. US10445643B2 · publicación de patente · 2019-10-15.

Describe una arquitectura de estados definidos por software. El documento no aporta evidencia específica de una auditoría de RAG con fecha de corte. Procedencia F21. Identificador, título y fecha de publicación o concesión comprobados. Ficha

B019 · Architectures, systems and methods for program defined transaction system and decentralized cryptocurrency system. US10679128B2 · publicación de patente · 2020-06-09.

Describe un sistema programable de transacciones y criptomonedas. Aporta contexto sobre registros distribuidos, sin demostrar un método de detección lingüística de información posterior a una fecha. Procedencia F22. Identificador, título y fecha de publicación o concesión comprobados. Ficha

B020 · Agentic Witnessing: Pragmatic and Scalable TEE-Enabled Privacy-Preserving Auditing. arXiv:2604.24203 · investigación · 2026-04-27.

Describe cómo un testigo puede atestiguar la actividad de un agente preservando su privacidad. Es una comparación posterior al corte sobre el alcance del entorno de ejecución de confianza. Procedencia F23, S3.4/S8. Identificador primario e historial de versiones comprobados. Ficha Otra fuente 1

B021 · AI Enforcement Accelerates as Federal Policy Stalls and States Step In - Morgan Lewis. comentario sobre actuaciones · sin fecha acreditada.

Análisis jurídico de la actividad supervisora federal y estatal. Para acreditar el inicio o resultado de una actuación concreta hay que consultar al organismo responsable. Procedencia F25. Página consultada. Sin verificación histórica. Ficha

B022 · Litigation Minute: Hype Responsibly: Legally Promoting Your AI | HUB | K&L Gates. comentario sobre actuaciones · sin fecha acreditada.

Comentario jurídico sobre publicidad de IA. Ayuda a interpretar el contexto, mientras que la constatación de un procedimiento requiere una fuente oficial. Procedencia F26. Pista registrada. Sin confirmación independiente del contenido o de la versión. Ficha

B023 · Interactive AI Whitepapers With Visual Diagrams | Veriprajna. pista de búsqueda · sin fecha acreditada.

Colección divulgativa de documentos y esquemas, incluida como vía de descubrimiento. La página actual no acredita cuándo se publicó por primera vez una propuesta. Procedencia F27. Pista registrada. Sin confirmación independiente del contenido o de la versión. Ficha

B024 · FTC Files New "AI Washing" Case | Perkins Coie. comentario sobre actuaciones · sin fecha acreditada.

Comentario jurídico sobre un caso de publicidad engañosa relacionada con IA. Para acreditar el procedimiento y sus términos se necesita el expediente de la FTC. Procedencia F28. Página consultada. Sin verificación histórica. Ficha

B025 · FTC's latest AI-washing case: A focus on agentic AI and productivity claims | DLA Piper. comentario sobre actuaciones · sin fecha acreditada.

Análisis de un caso sobre agentes de IA y promesas de productividad. Aporta contexto jurídico. Las actuaciones se comprueban en las fuentes del regulador. Procedencia F29. Pista registrada. Sin confirmación independiente del contenido o de la versión. Ficha

B026 · SEO vs GEO: Understanding the Key Differences in 2025 - Boomcycle Digital Marketing. contexto comercial · sin fecha acreditada.

Explicación comercial sobre visibilidad en buscadores convencionales y generativos. No compara la implementación de esas técnicas con el método de auditoría. Procedencia F30. Página consultada. Sin verificación histórica. Ficha

B027 · SEO vs GEO: What Is Generative Engine Optimization and Why It Matters in 2025. contexto comercial · sin fecha acreditada.

Recurso comercial sobre optimización de contenidos para motores generativos. Sirve como contexto editorial. No aporta una comparación técnica con la implementación auditada. Procedencia F31. Pista registrada. Sin confirmación independiente del contenido o de la versión. Ficha

B028 · GEO vs SEO: What's the Difference and Why It Matters in 2026 - DOJO AI. contexto comercial · sin fecha acreditada.

Artículo comercial sobre visibilidad en motores de búsqueda. No contiene una comparación de implementación con la auditoría temporal. Procedencia F32. Página consultada. Sin verificación histórica. Ficha

B029 · RAGPerf: An End-to-End Benchmarking Framework for Retrieval-Augmented Generation Systems. arXiv:2603.10765 · investigación · 2026-03-11.

Mide el rendimiento de sistemas RAG de extremo a extremo. Aporta contexto para diseñar evaluaciones, pero medir rendimiento no demuestra que se detecte contenido posterior a la fecha de corte. Procedencia F33. Identificador primario e historial de versiones comprobados. Ficha Otra fuente 1

B030 · Deterministic Legal Agents: A Canonical Primitive API for Auditable Reasoning over Temporal Knowledge Graphs. arXiv:2510.06002 · investigación · 2025-10-07.

Describe operaciones de recuperación y actuación sobre grafos jurídicos temporales. La v1 se titulaba Deterministic Legal Retrieval. Los cambios de formulación y las afirmaciones de versiones posteriores requieren sus propias fechas. Procedencia F34, F35. Identificador primario e historial de versiones comprobados. Ficha Otra fuente 1 Otra fuente 2

B031 · Architectures, systems and methods for program defined entertainment state system, decentralized cryptocurrency system and system with segregated secure functions and public functions. US20180247191A1 · publicación de patente · 2018-08-30.

Describe una arquitectura para estados de entretenimiento y criptomonedas con funciones protegidas separadas. La comparación de esta patente relacionada se limitó al identificador. Procedencia F36. Identificador, título y fecha de publicación o concesión comprobados. Ficha

B032–B072

B032 · SPE AUDIT — Cryptographic Integrity Verification for Digital …. descripción de proyecto · sin fecha acreditada.

Presentación de SPE AUDIT por su proveedor. Documenta la descripción del producto y no aporta validación independiente ni prueba de prioridad. Procedencia P01. Pista registrada. Sin confirmación independiente del contenido o de la versión. Ficha

B033 · About SPE Trace | How the Geo-Semantic Engine Works. descripción de proyecto · sin fecha acreditada.

Explicación de SPE Trace publicada por el propio proyecto. Describe cómo se presenta el producto al público. No acredita de forma independiente su rendimiento ni su prioridad. Procedencia P02, P03. Página consultada. Sin verificación histórica. Ficha Otra fuente 1

B034 · SPE TRACE — AI Brand Intelligence - Chrome Web Store. descripción de proyecto · sin fecha acreditada.

Ficha de SPE Trace en una tienda de extensiones. Documenta la presentación del producto en ese canal. De ella no se deducen validación independiente ni prioridad. Procedencia P04. Pista registrada. Sin confirmación independiente del contenido o de la versión. Ficha

B035 · 5 Best AI Audit Software Tools for Accounting Professionals - Vero AI. contexto comercial · sin fecha acreditada.

Comparativa comercial de herramientas de auditoría para profesionales contables. Se conserva como contexto de mercado, sin avalar el rendimiento de los proveedores. Procedencia P05. Página consultada. Sin verificación histórica. Ficha

B036 · Top 10 Best Audit AI Software of 2026 - Gitnux. contexto comercial · sin fecha acreditada.

Clasificación comercial de productos de auditoría con IA. Orienta sobre la oferta del mercado. No verifica el rendimiento anunciado de cada producto. Procedencia P06. Página consultada. Sin verificación histórica. Ficha

B037 · Best AI Evidence Check Platforms for Financial Audits | Finspectors. contexto comercial · sin fecha acreditada.

Comparativa sectorial de plataformas de verificación de evidencias financieras. Aporta contexto de mercado, sin validar las prestaciones de los proveedores. Procedencia P07. Página consultada. Sin verificación histórica. Ficha

B038 · The Evolving Role of Forensic Accountants in an AI-Driven World - Leonid Group. contexto comercial · sin fecha acreditada.

Comentario sectorial sobre la contabilidad forense y la IA. Sitúa el contexto profesional y no acredita el rendimiento de una herramienta. Procedencia P08. Pista registrada. Sin confirmación independiente del contenido o de la versión. Ficha

B039 · AI in the Public Sector: How Automation Is Transforming Government Finance - BDO USA. contexto comercial · sin fecha acreditada.

Comentario sectorial sobre automatización de finanzas públicas. Aporta contexto de mercado y no valida las prestaciones de productos de auditoría. Procedencia P09. Página consultada. Sin verificación histórica. Ficha

B040 · Top 6 AI Cyber Security Tools To Watch In 2026 - AccuKnox. contexto comercial · sin fecha acreditada.

Selección comercial de herramientas de ciberseguridad. Su utilidad se limita al contexto de mercado. No avala el rendimiento de los productos citados. Procedencia P10. Página consultada. Sin verificación histórica. Ficha

B041 · Top 10 Traceable AI Alternatives for API & App Security (2026) - Levo.ai. contexto comercial · sin fecha acreditada.

Comparativa comercial de productos de seguridad para aplicaciones e interfaces. Se conserva como referencia de mercado, sin validación del rendimiento. Procedencia P11. Página consultada. Sin verificación histórica. Ficha

B042 · The 10 Best AI Visibility Tools for 2026 - Evertune AI. contexto comercial · sin fecha acreditada.

Selección comercial de herramientas de visibilidad en IA. La presencia de un producto en la lista no acredita su rendimiento. Procedencia P12. Página consultada. Sin verificación histórica. Ficha

B043 · 9 Best Brand Monitoring Tools for AI Platforms in 2026. contexto comercial · sin fecha acreditada.

Comparativa comercial de seguimiento de marcas en plataformas de IA. Sirve como contexto de mercado. No constituye una evaluación independiente de las prestaciones. Procedencia P13. Página consultada. Sin verificación histórica. Ficha

B044 · 10 Best AI Brand Visibility Tools in (2026). contexto comercial · sin fecha acreditada.

Clasificación comercial de herramientas de visibilidad de marca. Ilustra la oferta del sector, sin verificar el rendimiento anunciado. Procedencia P14. Página consultada. Sin verificación histórica. Ficha

B045 · Top 10 Best LLM Brand Visibility Tools to Dominate AI Search in 2026 - Indexly. contexto comercial · sin fecha acreditada.

Comparativa comercial de herramientas que siguen la presencia de marcas en respuestas de modelos de lenguaje. Se usa para describir el mercado, sin avalar productos. Procedencia P15. Pista registrada. Sin confirmación independiente del contenido o de la versión. Ficha

B046 · The Guide to 9 Best AI Mode Tracking Tools in 2026 - SE Visible. contexto comercial · sin fecha acreditada.

Guía comercial de herramientas de seguimiento en modos de búsqueda con IA. Documenta la oferta del mercado. No acredita el rendimiento de las herramientas. Procedencia P16. Página consultada. Sin verificación histórica. Ficha

B047 · Semantic Density Effect (SDE): Maximizing Information Per Token Improves LLM Accuracy. arXiv:2604.17659 · investigación · 2026-04-19.

Estudia cómo cambia la precisión al variar la información por token de una instrucción. Aquí «densidad» designa una propiedad de la entrada. La incertidumbre de una respuesta y su respaldo documental fechado requieren otras medidas. Procedencia P17, P18, S3.4/S8, L2. Identificador primario e historial de versiones comprobados. Ficha Otra fuente 1 Otra fuente 2

B048 · Prompt-based justification and scoring with large language models for thought and language disorder assessment. 10.1007/s11042-026-21401-8 · investigación · 2026.

Estudia la puntuación de lenguaje clínico mediante instrucciones a un modelo. La relación con la auditoría temporal es indirecta. Este estudio no emplea datos clínicos. Procedencia P19. Fuente primaria pública identificada. Ficha Otra fuente 1

B049 · Youtu-VL: Unleashing Visual Potential via Unified Vision-Language Supervision. arXiv:2601.19798 · investigación · 2026-01-27.

Estudio de supervisión conjunta de visión y lenguaje. La cita no avala la afirmación independiente sobre JEPA que la acompañaba ni valida una auditoría temporal. Procedencia P20. Identificador primario e historial de versiones comprobados. Ficha Otra fuente 1

B050 · CausalRAG2: Hierarchical Causal Knowledge Graph Design for RAG. arXiv:2602.05143 · investigación · 2026-02-04.

La v1 se publicó como HugRAG y el trabajo se retituló CausalRAG2. Trata de recuperación con grafos causales jerárquicos. Es una obra distinta del CausalRAG publicado en marzo de 2025. Procedencia P21, P22, S3.4/S8, L2, Q0. Identificador primario e historial de versiones comprobados. Ficha Otra fuente 1 Otra fuente 2

B051 · Daily Papers - Hugging Face. pista de búsqueda · sin fecha acreditada.

Listado de novedades de investigación que ayuda a localizar artículos. La lista no demuestra por sí sola la existencia histórica de un mecanismo con unas características concretas. Procedencia P23. Página consultada. Sin verificación histórica. Ficha

B052 · CausalRAG: Integrating Causal Graphs into Retrieval-Augmented Generation. arXiv:2503.19878 · investigación · 2025-03-25.

CausalRAG emplea grafos causales en la recuperación. Su v1 de marzo de 2025 precede al corte. Véase M39. La presencia de un grafo causal en la búsqueda no demuestra el origen de una respuesta. Procedencia P24, L2. Identificador primario e historial de versiones comprobados. Ficha Otra fuente 1

B053 · Jing Ma - CatalyzeX. pista de búsqueda · sin fecha acreditada.

Perfil de Jing Ma usado para localizar publicaciones. Por sí solo no documenta la implementación histórica de un mecanismo. Procedencia P25. Pista registrada. Sin confirmación independiente del contenido o de la versión. Ficha

B054 · Professor Ah-Lian Kor | Leeds Beckett University. pista de búsqueda · sin fecha acreditada.

Perfil académico de Ah-Lian Kor usado como vía de descubrimiento. La existencia del perfil no acredita el alcance técnico ni la primera divulgación de un trabajo. Procedencia P26. Página consultada. Sin verificación histórica. Ficha

B055 · Daily Papers - Hugging Face. pista de búsqueda · sin fecha acreditada.

Índice de novedades de Hugging Face. Permite buscar obras. Para acreditar históricamente un mecanismo hay que consultar la obra y su versión. Procedencia P27. Pista registrada. Sin confirmación independiente del contenido o de la versión. Ficha

B056 · Key Issue 5: Transparency Obligations - EU AI Act. comentario normativo · sin fecha acreditada.

Interpretación de terceros sobre transparencia. La correspondencia con artículos del Reglamento de IA de la UE se basa en el texto legislativo consolidado. Procedencia P28. Página consultada. Sin verificación histórica. Ficha

B057 · High-level summary of the AI Act | EU Artificial Intelligence Act. comentario normativo · sin fecha acreditada.

Resumen del Reglamento de IA elaborado por terceros. Para relacionar controles técnicos y artículos, el estudio se remite al texto legislativo consolidado. Procedencia P29. Página consultada. Sin verificación histórica. Ficha

B058 · Article 12: Record-Keeping | EU Artificial Intelligence Act. comentario normativo · sin fecha acreditada.

Explicación de terceros del artículo 12 sobre conservación de registros. El análisis del artículo en este estudio se basa en el texto consolidado del Reglamento. Procedencia P30. Página consultada. Sin verificación histórica. Ficha

B059 · Iso 42001 Logging Lifecycle Traceability Vs Eu Ai Act - ISMS.online. comentario normativo · sin fecha acreditada.

Comparación de terceros entre gestión de registros según ISO 42001 y el Reglamento de IA. Las obligaciones por artículo se contrastan con el texto legislativo consolidado. Procedencia P31. Página consultada. Sin verificación histórica. Ficha

B060 · Operationalizing Compliance with the EU AI Act's High-Risk Requirements - Label Studio. comentario normativo · sin fecha acreditada.

Guía de terceros para aplicar requisitos de sistemas de alto riesgo. La correspondencia jurídica empleada en el estudio se establece a partir del texto consolidado del Reglamento. Procedencia P32. Página consultada. Sin verificación histórica. Ficha

B061 · Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1 · contexto institucional · 2023-01-26.

Marco de gestión de riesgos de IA de NIST. Para interpretar una fuente de política pública hay que identificar su emisor y alcance. Un comentario remitido en una consulta carece por sí mismo de la autoridad de una guía de la agencia. Procedencia P33. Edición primaria fechada del programa NIST citado. Ficha Otra fuente 1

B062 · Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1 · guidance · 2024-07.

Guía de NIST sobre riesgos de IA generativa. Aporta contexto de gestión de riesgos. No establece una prueba de rendimiento ni una determinación jurídica. Procedencia P34, P35. Fuente primaria pública identificada. Ficha Otra fuente 1 Otra fuente 2

B063 · NIST AI RMF Implementation Guide (2026) + EU AI Act Map | Knowlee. comentario normativo · sin fecha acreditada.

Interpretación de terceros sobre NIST AI RMF y el Reglamento de IA. Las obligaciones europeas se cotejan por artículo en el texto legislativo consolidado. Procedencia P36. Página consultada. Sin verificación histórica. Ficha

B064 · Safeguard the Future of AI: The Core Functions of the NIST AI RMF - Optro. comentario normativo · sin fecha acreditada.

Explicación de terceros de las funciones del marco NIST AI RMF. La correspondencia regulatoria europea del estudio se basa en el texto legislativo consolidado. Procedencia P37. Página consultada. Sin verificación histórica. Ficha

B065 · Treasury Launches the Artificial Intelligence (AI) Innovation Series. contexto institucional · sin fecha acreditada.

Anuncio del Tesoro estadounidense sobre una iniciativa de innovación en IA. Documenta una actuación de política pública. Conviene distinguir los documentos oficiales de los comentarios remitidos por terceros. Procedencia P38. Página consultada. Sin verificación histórica. Ficha

B066 · Re: Request for Comment on Innovative Methods to Detect Illicit Activity Involving Digital Assets (Docket No. TREAS-DO-2025-0009 - SEC.gov. contexto institucional · sin fecha acreditada.

Escrito presentado en respuesta a una consulta pública. Atribuye la posición a quien la presenta. Su incorporación al expediente no la convierte en orientación oficial del organismo. Procedencia P39. Página consultada. Sin verificación histórica. Ficha

B067 · United States Patent and Trademark Office Fiscal Year 2027 Congressional Submission April 2026 - USPTO. contexto administrativo · sin fecha acreditada.

Documento presupuestario de la USPTO. Ayuda a entender el contexto económico de la revisión. Los importes pueden cambiar y no constituyen un antecedente técnico. Procedencia P40. Página consultada. Sin verificación histórica. Ficha

B068 · USPTO Fee Schedule - Current. contexto administrativo · sin fecha acreditada.

Tabla de tasas de la USPTO. Los importes pueden cambiar y no aportan antecedentes técnicos. La variante del nombre de archivo P46 se agrupó por organismo emisor y denominación. No se verificaron la identidad de bytes ni su revisión histórica. Procedencia P41, P46. Página consultada. Sin verificación histórica. Ficha Otra fuente 1

B069 · Table of Patent Fees - Current, Final Rule and Unit Cost - USPTO. contexto administrativo · sin fecha acreditada.

Tabla comparativa de tasas y costes de la USPTO. Aporta contexto económico. Las tarifas pueden cambiar y no acreditan una divulgación técnica. Procedencia P42. Página consultada. Sin verificación histórica. Ficha

B070 · Table of Patent Fees - USPTO. contexto administrativo · sin fecha acreditada.

Tabla de tasas de la USPTO. Se conserva como contexto administrativo. Los importes están sujetos a cambios y carecen de valor como antecedente técnico. Procedencia P43. Pista registrada. Sin confirmación independiente del contenido o de la versión. Ficha

B071 · USPTO fee schedule. contexto administrativo · sin fecha acreditada.

Tarifario de la USPTO. El importe aplicable depende de la fecha. Se documentó por su relevancia para el coste de la revisión, sin atribuirle contenido técnico. Procedencia P44. Página consultada. Sin verificación histórica. Ficha

B072 · USPTO Fee Changes to Take Effect in January 2025: What You Need to Know - Honigman. contexto administrativo · sin fecha acreditada.

Comentario profesional sobre los cambios de tasas previstos para enero de 2025. Es contexto presupuestario sujeto a revisión y no aporta antecedentes técnicos. Procedencia P45. Página consultada. Sin verificación histórica. Ficha

B073–B094

B073 · ZKPROV: A Zero-Knowledge Approach to Dataset Provenance for Large Language Models. arXiv:2506.20915 · investigación · 2025-06-26.

ZKPROV aborda la prueba de procedencia de conjuntos de datos para modelos de lenguaje. Su v1 de junio de 2025 es anterior al corte, aunque figurase entre los trabajos posteriores. La v2 de diciembre tiene fecha propia. Procedencia S3.4/S8. Identificador primario e historial de versiones comprobados. Ficha

B074 · VeriLLM: A Lightweight Framework for Publicly Verifiable Decentralized Inference. arXiv:2509.24257 · investigación · 2025-09-29.

VeriLLM estudia la inferencia descentralizada verificable por el público. El primer manuscrito es de septiembre de 2025. Las afirmaciones introducidas en revisiones posteriores requieren fechas propias. Procedencia S3.4/S8. Identificador primario e historial de versiones comprobados. Ficha

B075 · MathLedger: A Verifiable Learning Substrate with Ledger-Attested Feedback. arXiv:2601.00816 · investigación · 2025-12-22.

MathLedger registra y atestigua la realimentación usada durante el aprendizaje. Se envió el 22 de diciembre de 2025, aunque el identificador de arXiv corresponda a enero. Es una comparación posterior al corte. Procedencia S3.4/S8. Identificador primario e historial de versiones comprobados. Ficha

B076 · SecureInfer: Heterogeneous TEE-GPU Architecture for Privacy-Critical Tensors for Large Language Model Deployment. arXiv:2510.19979 · investigación · 2025-10-22.

SecureInfer separa el procesamiento de tensores sensibles en un entorno protegido y la aceleración en GPU. Esta distribución delimita qué componentes deben considerarse de confianza. La ejecución confidencial por sí sola no acredita una afirmación temporal. Procedencia T-B. Identificador primario e historial de versiones comprobados. Ficha

B077 · TZ-LLM: Protecting On-Device Large Language Models with Arm TrustZone. arXiv:2511.13717 · investigación · 2025-11-17.

TZ-LLM protege modelos ejecutados en el propio dispositivo mediante Arm TrustZone. Aporta mecanismos de integridad y confidencialidad de ejecución. La fecha de un hecho requiere verificación independiente. Procedencia T-B. Identificador primario e historial de versiones comprobados. Ficha

B078 · AgenTEE: Confidential LLM Agent Execution on Edge Devices. arXiv:2604.18231 · investigación · 2026-04-20.

AgenTEE estudia la ejecución confidencial de agentes en el propio dispositivo. Es una referencia posterior al corte para comparar mecanismos de aislamiento. Procedencia T-B. Identificador primario e historial de versiones comprobados. Ficha

B079 · FlexServe: A Fast and Secure LLM Serving System for Mobile Devices with Flexible Resource Isolation. arXiv:2603.09046 · investigación · 2026-03-10.

FlexServe ofrece un servicio de inferencia móvil con aislamiento flexible de recursos. Es una comparación posterior. El trabajo no estaba disponible en la fecha de referencia de diciembre. Procedencia T-B. Identificador primario e historial de versiones comprobados. Ficha

B080 · Protecting Context and Prompts: Deterministic Security for Non-Deterministic AI. arXiv:2602.10481 · investigación · 2026-02-11.

Estudia la integridad autenticada del contexto y de la instrucción. Es un precedente posterior para la integridad de entradas. Las afirmaciones sobre la veracidad histórica del texto suministrado necesitan evidencia independiente. Procedencia T-B. Identificador primario e historial de versiones comprobados. Ficha

B081 · SEAL: Entangled White-box Watermarks on Low-Rank Adaptation. arXiv:2501.09284 · investigación · 2025-01-16.

SEAL incorpora marcas de agua en adaptaciones LoRA para respaldar atribuciones de titularidad del modelo. Fijar criptográficamente la versión utilizada durante una ejecución requiere una comprobación distinta. Procedencia T-B. Identificador primario e historial de versiones comprobados. Ficha

B082 · Cloud storage using Merkle trees. US9705730B1 · publicación de patente · 2017-07-11.

Describe almacenamiento en la nube y verificación de integridad con árboles de Merkle. Un compromiso genérico de datos no acredita que un hecho fuese desconocido antes de una fecha. Procedencia T-B4. Identificador, título y fecha de publicación o concesión comprobados. Ficha

B083 · Method and system for verifying information of a data item in a plurality of different data items . US20180026791A1 · publicación de patente · 2018-01-25.

Describe la verificación de información seleccionada ligada a un elemento de datos. Permite comparar mecanismos que autentican los elementos de un conjunto. Procedencia T-B4. Identificador, título y fecha de publicación o concesión comprobados. Ficha

B084 · Platform for multi-party digital records using distributed ledger system. US20190253240A1 · publicación de patente · 2019-08-15.

Plataforma de registros digitales con varias partes mediante un registro distribuido. La comparación no incluye un análisis de infracción reivindicación por reivindicación. Procedencia T-B4. Identificador, título y fecha de publicación o concesión comprobados. Ficha

B085 · Methods and devices for validating data in a blockchain network. US20230237477A1 · publicación de patente · 2023-07-27.

Describe la validación de datos en una red de cadena de bloques. Es una referencia de infraestructura de verificación. No contiene una prueba de evaluación de texto histórico. Procedencia T-B4. Identificador, título y fecha de publicación o concesión comprobados. Ficha

B086 · Merkle proof entity. US20230394063A1 · publicación de patente · 2023-12-07.

Describe pruebas de Merkle y las entidades asociadas. Acreditar la pertenencia a un conjunto demuestra una propiedad más limitada que determinar qué sabía un modelo. Procedencia T-B4. Identificador, título y fecha de publicación o concesión comprobados. Ficha

B087 · KEA Explain: Explanations of Hallucinations using Graph Kernel Analysis. arXiv:2507.03847 · investigación · 2025-07-05.

KEA Explain usa núcleos sobre grafos para explicar alucinaciones. Es pertinente para comparar significado y respaldo. Su puntuación no atestigua la fecha de la evidencia. Procedencia Q-G, Q-T. Identificador primario e historial de versiones comprobados. Ficha Otra fuente 1

B088 · Detecting hallucinations in large language models using semantic entropy. 10.1038/s41586-024-07421-0 · investigación · 2024-06-19.

La entropía sobre grupos de respuestas semánticamente equivalentes puede detectar ciertas invenciones. La confianza estimada no permite por sí sola fechar la divulgación de un hecho. Procedencia F-body, Q0. Fuente primaria pública identificada. Ficha

B089 · Jolt: SNARKs for Virtual Machines via Lookups. IACR:2023/1217 · investigación · 2023.

Jolt aborda la computación verificable de propósito general. Una prueba sobre una entrada no establece la disponibilidad histórica de los hechos descritos en ella. Procedencia Q-G. Fuente primaria pública identificada. Ficha

B090 · vCNN: Verifiable Convolutional Neural Network Based on zk-SNARKs. IACR:2020/584 · investigación · 2020.

vCNN verifica cálculos de redes neuronales convolucionales. Esa verificación no documenta cuándo estuvo disponible una información para el modelo. Procedencia Q-G. Fuente primaria pública identificada. Ficha

B091 · Neural Authenticity Ledger: Cryptographic Model State Fingerprinting for AI Content Provenance. NAL · self-published proposal · 2025-09-20.

Propuesta autopublicada que vincula huellas del estado del modelo con la procedencia de contenidos. La fecha procede del editor. No se acreditaron un depósito independiente, efectos jurídicos ni una implementación. Procedencia T-B3. Página y fecha declaradas por su editor. Ficha

B092 · V-Proof Protocol. V-Proof · proyecto sectorial · sin fecha acreditada.

Un directorio público de empresas identifica el producto V-Proof Protocol. La ficha no acredita una divulgación suficientemente detallada para reproducirlo ni la primera fecha de acceso público. Procedencia T-B4. Entrada identificada en el directorio. Ficha

B093 · GRAIL Protocol (historical label). GRAIL · denominación sin identificar · sin fecha acreditada.

Varios proyectos independientes emplean el nombre GRAIL Protocol. No se pudo identificar inequívocamente el que correspondería a la descripción de inferencia verificable. Procedencia S3.4/S8. Identidad sin resolver. Excluida de la comparación técnica.

B094 · MECM (temporal causal gating, historical label). MECM · denominación sin identificar · sin fecha acreditada.

La pista MECM ofrecía una sigla, sin identificador único, título ni autor. Al no poder identificar la fuente, quedó fuera de las conclusiones técnicas. Procedencia Q-G, Q-T. Identidad sin resolver. Excluida de la comparación técnica.

B095–B109

B095 · Giza. Giza · proyecto sectorial · sin fecha acreditada.

Página del propio proyecto Giza, perteneciente al ámbito del aprendizaje automático verificable. La página cambia con el tiempo y no permite establecer qué funciones existían al corte. Procedencia Q-C. Página del proyecto sin versión fechada comprobada. Ficha

B096 · Aleo. Aleo · proyecto sectorial · sin fecha acreditada.

Página del propio proyecto Aleo para localizar su infraestructura de ejecución verificable. No quedó fijada una versión histórica concreta. Procedencia Q-C. Página del proyecto sin versión fechada comprobada. Ficha

B097 · Internet X.509 Public Key Infrastructure Time-Stamp Protocol (TSP). RFC 3161 · norma · 2001-08.

Una autoridad de sellado temporal puede vincular una huella digital a la hora declarada, bajo los supuestos de confianza del servicio. El sello no determina cuándo ocurrió el hecho descrito en los datos. Procedencia D2/D4. Fuente primaria pública identificada. Ficha

B098 · Information technology — Artificial intelligence — Management system. ISO/IEC 42001:2023 · norma · 2023.

Norma de sistemas de gestión de IA mencionada en relación con las obligaciones de registro. La cita no constituye una certificación. Procedencia P-body. Registro identificado en el catálogo ISO. Ficha

B099 · Artificial Intelligence Act. Regulation (EU) 2024/1689 · legislation · 2024-07-12.

Marco regulador citado en ambas revisiones. Cada obligación debe cotejarse con el texto consolidado aplicable y el calendario de aplicación del artículo correspondiente. La evidencia técnica por sí sola no acredita conformidad. Procedencia F/P-body. Fuente primaria pública identificada. Ficha Otra fuente 1

B100 · SEC charges two investment advisers with making false and misleading statements about their use of artificial intelligence. SEC-2024-36 · actuación oficial · 2024-03-18.

Comunicado de la SEC sobre dos asesores de inversión, citado para identificar una actuación relacionada con publicidad y divulgación de información. Las resoluciones no validan el método estudiado. Procedencia F-body. Fuente primaria pública identificada. Ficha

B101 · FTC Announces Crackdown on Deceptive AI Claims and Schemes. FTC-2024-09-AI · actuación oficial · 2024-09-25.

Anuncio oficial de la FTC sobre varios procedimientos. Acredita que la agencia hizo el anuncio y permite reemplazar los resúmenes de terceros para documentar ese hecho. Procedencia F-body. Fuente primaria pública identificada. Ficha

B102 · Alice Corp. v. CLS Bank International. USSC-13-298 · resolución judicial · 2014.

Alice se cita por su papel en el debate estadounidense sobre qué materias pueden patentarse. La sentencia no se trata como divulgación técnica ni determina la patentabilidad de este expediente. Procedencia L1. Resolución identificada. Ficha

B103 · Enfish, LLC v. Microsoft Corp.. CAFC-15-1244 · resolución judicial · 2016.

Enfish se cita como jurisprudencia estadounidense sobre materia patentable. De la decisión no se deriva un antecedente técnico ni una conclusión de patentabilidad para este expediente. Procedencia L1. Resolución identificada. Ficha

B104 · McRO, Inc. v. Bandai Namco Games America Inc.. CAFC-15-1080 · resolución judicial · 2016.

McRO aporta contexto jurisprudencial estadounidense sobre materia patentable. No se le atribuyen una divulgación técnica ni una decisión sobre la patentabilidad del método aquí estudiado. Procedencia L1. Resolución identificada. Ficha

B105 · BASCOM Global Internet Services, Inc. v. AT&T Mobility LLC. CAFC-15-1763 · resolución judicial · 2016.

BASCOM figura como contexto estadounidense sobre materia patentable. Su inclusión no la convierte en un antecedente técnico ni establece la patentabilidad de este expediente. Procedencia L1. Resolución identificada. Ficha

B106 · Berkheimer v. HP Inc.. CAFC-17-1437 · resolución judicial · 2018.

Berkheimer se conserva como referencia judicial estadounidense sobre materia patentable. No se extrae de la sentencia una divulgación técnica ni una conclusión específica sobre la solicitud estudiada. Procedencia L1. Resolución identificada. Ficha

B107 · Diamond v. Diehr. US-450-175 · resolución judicial · 1981.

Diamond v. Diehr se cita como contexto jurisprudencial estadounidense sobre materia patentable. La sentencia no se invoca como divulgación técnica ni como resolución de este expediente. Procedencia L1. Resolución identificada. Ficha

B108 · Finjan, Inc. v. Blue Coat Systems, Inc.. CAFC-16-2520 · resolución judicial · 2018.

Finjan ilustra el debate estadounidense sobre materia patentable. Aquí no se le atribuyen una divulgación técnica ni una determinación de patentabilidad para el método estudiado. Procedencia L1. Resolución identificada. Ficha

B109 · Core Wireless Licensing S.A.R.L. v. LG Electronics, Inc.. CAFC-16-2684 · resolución judicial · 2018.

Core Wireless se cita por la discusión estadounidense sobre materia patentable. No aporta una divulgación técnica de este método ni resuelve la patentabilidad del expediente. Procedencia L1. Resolución identificada. Ficha

B110–B116

B110 · ASN.1 encoding rules, including Distinguished Encoding Rules (DER). X.690 · norma · 2021-02.

La especificación identifica las reglas de codificación ASN.1, incluidas las DER, mencionadas en la revisión. Normalizar los bytes permite reproducir una codificación. No identifica toda la evidencia disponible en una fecha. Procedencia F-body. Catálogo de la UIT, edición de febrero de 2021. Ficha

B111 · Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs. arXiv:1603.09320 · investigación · 2016-03-30.

Artículo original sobre HNSW, método citado en la revisión. Una búsqueda aproximada de vecinos más próximos puede omitir elementos admisibles. Esas omisiones importan al interpretar el resultado de la auditoría. Procedencia F-body. Identificador primario comprobado. Ficha

B112 · SEC proceeding concerning Presto Automation. SEC-33-11352 · actuación oficial · 2025-01-14.

La SEC anunció la resolución acordada de cargos relacionados con declaraciones sobre Presto Voice. Una descripción anterior había atribuido erróneamente el procedimiento a la FTC. Procedencia F-body. Fuente primaria pública identificada. Ficha

B113 · SEC civil proceedings concerning representations about Nate. SEC-LR-26282 · actuación oficial · 2025-04-11.

Según la SEC, su demanda civil en el caso Nate se presentó el 9 de abril de 2025. La fecha consignada en la ficha corresponde al comunicado. Procedencia F-body. Fuente primaria pública identificada. Ficha

B114 · Department of Justice announcement concerning Nate. SDNY-nate-2025-04-09 · actuación oficial · 2025-04-09.

Anuncio del Departamento de Justicia sobre una acusación penal en el caso Nate. Es un procedimiento distinto de la demanda civil de la SEC. El anuncio recoge alegaciones. Procedencia F-body. Fuente primaria pública identificada. Ficha

B115 · FTC proceedings concerning Air AI. FTC-AirAI · actuación oficial · 2025-08-25.

La fecha corresponde a la demanda de la FTC sobre Air AI. La actuación abarca la comercialización, las afirmaciones de ingresos y los reembolsos. La misma página documenta también un acuerdo posterior. Procedencia F-body. Fuente primaria pública identificada. Ficha

B116 · The New York Times Company v. Microsoft Corp. et al., complaint. SDNY-1:23-cv-11195 · court pleading · 2023-12-27.

Copia pública de la demanda presentada por The New York Times, citada para identificar un litigio de derechos de autor. Las alegaciones del escrito no constituyen hechos declarados probados. Procedencia F-body. Demanda fechada identificada. Ficha