Auditoría temporal de las respuestas de IA

Rubén Abella · SPE Audit
Edición de investigación 1.2 · 22 de noviembre de 2026

Resumen

Cuatro sucesos de misiones espaciales sustentan la comparación original. Las 96 respuestas de desarrollo combinan cuatro sucesos, ocho formulaciones y tres fechas de corte. Otras 16 respuestas cambian la redacción y conservan los mismos sucesos. Examinamos si una frase da por ocurrido un desenlace posterior al corte y qué evidencia puede conservar una aplicación para que otra persona revise esa decisión.

La regla que conocía la fecha del desenlace detectó 24 de 32 positivos en desarrollo. Al cambiar la redacción, detectó 0 de 8 y remitió 14 de 16 casos a revisión. Sus dos decisiones automáticas fueron correctas. Una persona que recibiera las 16 respuestas seguiría teniendo que examinar las otras catorce.

En esta edición, un modelo local evaluó las mismas 112 respuestas con y sin un campo que identifica la fecha del desenlace. Añadirlo cambió seis decisiones de desarrollo y ninguna del reto. En el reto, ambos pases detectaron **4/8** positivos, emitieron una falsa alerta y remitieron **11/16** casos.

El autor desarrolla la tecnología evaluada. El proyecto utilizó asistencia de IA en la preparación de textos, reglas, análisis y documentación. Las referencias de clasificación son editoriales. El piloto de revisión humana sigue abierto.

DOI de esta edición · 10.5281/zenodo.24447732

Qué se afirma y qué queda registrado

Antes del impacto de DART ya era público que la misión intentaría chocar con Dimorphos. A fecha de 1 de enero de 2022, «está previsto un impacto» expresa un plan. «DART impactó» presenta el desenlace como consumado. La NASA comunicó el impacto del 26 de septiembre de 2022. La diferencia depende del verbo, del referente y del corte, incluso si ambas frases contienen el mismo año.

Término	Criterio en este estudio
Positivo de referencia	La respuesta afirma que el suceso posterior al corte ya ocurrió.
Control de referencia	La respuesta expresa otra proposición, como un plan, una negación o un hecho anterior al corte.
Alerta / Otro / Revisión	Salidas del método. «Revisión» aplaza la decisión. Una referencia indeterminada expresa ambigüedad en la propia etiqueta.

Hasta dónde llega la captura

El adaptador registra los bytes de la petición entregada al emisor local, el texto devuelto y sus huellas. Un revisor puede recuperar la pregunta, el corte y las fichas admitidas y comprobar si coinciden con un registro conservado.

Supóngase que dos servicios devuelven la misma frase. Uno reproduce una previsión acertada. El otro consulta una crónica posterior fuera de la aplicación. Si el cliente observa los mismos bytes en ambos casos, también obtiene las mismas huellas. Para atribuir la respuesta a una fuente haría falta observar algo que distinga esos dos recorridos.

En el estudio original, **124 capturas** reúnen 96 respuestas de desarrollo, 16 de redacción nueva y 12 pruebas del selector local. En estas últimas, una función devolvía el texto de la ficha admisible más reciente. Retirar la ficha comprobaba el comportamiento de esa función. Los resultados de clasificación emplean solo las primeras 112 respuestas. El editor del reto de 16 conocía la pregunta de investigación y el límite general de la regla anual. Redactó los textos con asistencia de IA antes de consultar las reglas fechadas o los ejemplos de desarrollo.

El ensayo con y sin la fecha del desenlace

Para cada respuesta ejecutamos el mismo lector local dos veces. Ambos pases reciben el corte inclusivo, la pregunta, la respuesta y, en su orden registrado, la fecha y el texto de cada ficha admisible. Se retiran las URL y los identificadores descriptivos de las fichas. Un pase añade un único campo con la fecha registrada del desenlace, que puede ser posterior al corte. El otro lo omite.

Método	Información disponible para decidir
Regla anual 1.0	Años explícitos de la respuesta y año del corte. Se mantiene desactivada la señal de contexto futuro del ensayo original.
Regla fechada 1.0	Texto, corte, fecha registrada del suceso y expresiones de desenlace elegidas para cada misión. Puede abstenerse.
Modelo local 1.2	Mismo modelo y mismas entradas en ambos pases salvo el campo de fecha. La plantilla permite alerta, otro o revisión.

Modelo y ejecución. SmolLM2-135M-Instruct, pesos Q4_K_M, llama.cpp b8058, CPU, temperatura 0 y semilla 20261122. Cada petición empieza sin historial de conversación. El orden de los dos pases se alterna por fila. Son 224 inferencias sobre respuestas ya escritas. Las etiquetas se unen por identificador al puntuar. Las respuestas originales no proceden de esas inferencias.

El corpus y sus resultados previos ya eran públicos. Antes de la nueva ejecución se fijaron la instrucción, las entradas, el código y sus huellas. DigiCert emitió un sello RFC 3161 sobre la huella del plan el **22 de noviembre de 2026 a las 19:23:03 UTC**. Ese sello fecha el compromiso con el plan. Las etiquetas siguen procediendo del proyecto, y las fechas presentes en las fichas, el texto o el entrenamiento del modelo aún pueden orientar el pase que omite el campo.

`SHA-256 9402541a361800c9b2f7dddf1ccdb36c8cd6e383c7b46c6a5838f34c0252073e`

Plan previo · Instrucción exacta y sello verificable

Decisiones y carga de revisión

Las fracciones conservan su denominador. **Detectados** son alertas correctas sobre todos los positivos. **Falsas** son alertas sobre todos los controles. **Precisión** es la proporción de alertas correctas. **Cobertura** cuenta las decisiones automáticas sobre todas las filas. **Revisión** cuenta las abstenciones sobre esas mismas filas.

96 respuestas de desarrollo · 32 positivos, 64 controles

Método	Detectados	Falsas	Precisión	Cobertura	Revisión
Regla anual	4/32	12/64	4/16	96/96	0/96
Regla fechada	24/32	0/64	24/24	84/96	12/96
Modelo sin fecha	8/32	16/64	8/24	55/96	41/96
Modelo con fecha	8/32	17/64	8/25	54/96	42/96

16 respuestas de redacción nueva · 8 positivos, 8 controles

Método	Detectados	Falsas	Precisión	Cobertura	Revisión
Regla anual	1/8	2/8	1/3	16/16	0/16
Regla fechada	0/8	0/8	—	2/16	14/16
Modelo sin fecha	4/8	1/8	4/5	5/16	11/16
Modelo con fecha	4/8	1/8	4/5	5/16	11/16

Aciertos entre decisiones automáticas, en el orden de las tablas: 56/96, 84/84, 29/55 y 27/54 en desarrollo. En el reto: 7/16, 2/2, 4/5 y 4/5. Los 224 recibos del modelo aportaron una etiqueta válida. La precisión de la regla fechada en el reto queda sin definir porque no emitió ninguna alerta.

La diferencia entre el 12,5 % de sensibilidad de la regla anual y el 75 % de la regla fechada en desarrollo combina cambios de método y de información disponible. En el reto, acertar 2/2 entre las decisiones automáticas de la regla fechada deja una cobertura de 2/16. Los ocho positivos quedaron pendientes. Ninguna de estas fracciones estima una tasa de error en producción.

Decisiones por fila · Recuentos por suceso · Transiciones por etiqueta

Por qué conviene mirar cada misión

En las 96 respuestas de desarrollo, ambos pases del modelo detectaron ocho de los 32 positivos. Los ocho pertenecían a Perseverance. El modelo emitió una alerta en cada una de las 24 respuestas de esa misión, incluidas las 16 de control. En las otras tres misiones dejó pasar o remitió todos los positivos de desarrollo.

Misión	Positivos detectados sin / con fecha	Falsas alertas sin / con fecha	A revisión sin / con fecha
Perseverance	8/8 · 8/8	16/16 · 16/16	0/24 · 0/24
DART	0/8 · 0/8	0/16 · 1/16	17/24 · 18/24
OSIRIS-REx	0/8 · 0/8	0/16 · 0/16	0/24 · 0/24
Rosetta	0/8 · 0/8	0/16 · 0/16	24/24 · 24/24

Las seis decisiones que cambiaron pertenecían a controles de DART. Tres pasaron de «otro» a revisión, dos de revisión a «otro» y uno de «otro» a alerta. Por tanto, el mismo número de positivos detectados puede coexistir con una falsa alarma más y con distintos expedientes en la cola de revisión.

En el reto de 16 textos, ambos pases tomaron las mismas 16 decisiones. Los cuatro positivos detectados se reparten entre Perseverance y DART, dos en cada misión. Rosetta y OSIRIS-REx aportan ocho respuestas al reto. El modelo remitió las ocho a revisión. Cada misión aporta solo dos positivos en este conjunto.

Cifras por misión y método · Decisiones emparejadas

Ocho decisiones que se pueden inspeccionar

Se conservan los ocho ejemplos seleccionados para el artículo original. Las dos últimas columnas añaden las lecturas medidas en 1.2. **P** = positivo de referencia o alerta del detector, **O** = otro y **R** = revisión. «Sin» y «con» indican la ausencia o presencia del campo de fecha del desenlace en el modelo local. Las salidas se obtuvieron con los textos en inglés.

Caso	Lectura de la frase	Ref.	Anual	Fechada	Sin	Con
perseverance-04	Prevé el aterrizaje	O	P	O	P	P
perseverance-05	Niega la confirmación	O	P	O	P	P
osiris-06	Cita una pregunta	O	P	O	O	O
rosetta-08	Afirma el despertar	P	O	P	R	R
dart-09	Afirma el impacto	P	O	P	R	R
h03	Plantea una hipótesis	O	O	O	R	R
h10	Afirma un hecho anterior	O	O	R	R	R
h16	Afirma «lo hizo»	P	O	R	P	P

Cuando un pronombre cambia la lectura

«Perseverance was built to land on Mars. It did so in February 2021.» El corte de **h16** es el 10 de febrero. La primera frase explica para qué se construyó el róver. La segunda retoma el verbo y afirma el aterrizaje, que ocurrió el día 18. La regla fechada remitió el caso a revisión. En cambio, la negación «There is no confirmation that Perseverance landed on Mars in 2021» es un control ante el corte de 2020. La regla anual emitió una falsa alerta ante ese año futuro.

Los ocho ejemplos ilustran desacuerdos concretos. Quien quiera estimar cuántas veces sucedió cada cosa dentro del ensayo puede incluir las 112 filas del fichero público.

Textos de desarrollo · Textos del reto

Qué se observó fuera de las cuatro misiones

Seis sucesos nuevos, 48 respuestas en español e inglés

El suplemento 1.1 evaluó 18 positivos, 24 controles y seis referencias indeterminadas de transporte, artes escénicas y física experimental. La regla de proposiciones detectó **1/18**, emitió una falsa alerta entre 24 controles y remitió **24/48** a revisión. En el caso v038, la respuesta cita la premisa de una inauguración y después la rechaza. La regla se quedó con la afirmación de la primera cláusula.

La lectura exploratoria asistida por IA coincidió con los 18 positivos, los 24 controles y las seis ambigüedades. Se encargó una vez conocidos los resultados agregados de la regla. Un seguimiento posterior empleó un lector local con instrucción y pesos fijados. Detectó **11/18** positivos, alertó en **13/24** controles y remitió **17/48** casos. Sus decisiones se repitieron en un segundo pase. Las dos lecturas tienen distinto protocolo y distinto valor como evidencia.

Diecisiete pasajes escritos por terceros

Seis sucesos del NIST y el USGS aportaron 17 pasajes en inglés. Las instituciones publicaron las frases. El proyecto eligió los pasajes, asignó cortes históricos y redactó las preguntas y etiquetas. El lector local del seguimiento detectó **0/5** positivos y remitió **16/17** pasajes. Su única decisión automática cerró como «otro» un positivo.

La selección dirigida aporta otro origen para la redacción. Las referencias continúan siendo internas al proyecto. El ejercicio de traducción a cuatro idiomas reutiliza las mismas respuestas originales. Ninguno de estos conjuntos permite deducir la frecuencia de afirmaciones posteriores al corte en un servicio remoto.

Suplemento 1.1 · Resultados del seguimiento y fuentes institucionales

Antecedentes y alcance de la evidencia

Cuatro trabajos ayudan a situar la aportación. La revisión se dirigió a operaciones técnicas concretas y versiones identificables. El catálogo adjunto permite seguir también los duplicados, las búsquedas y las pistas cuyo origen continúa abierto.

Antecedente	Qué aporta a esta pregunta
Dantart · N01	Recuperación histórica y registro firmado por consulta, con documentos, respuesta y política temporal.
EQTY Lab · N02	Evidencia criptográfica que enlaza datos recuperados con una entrada de inferencia. El punto atestado delimita qué se acredita.
Liu y cols., ExAnte · N03	Referencias anteriores y posteriores para evaluar afirmaciones cuya verificación depende de la fecha.
Paleka y cols. · N06	Páginas antiguas actualizadas con el desenlace y sesgos de búsqueda que pueden sobrevivir al filtro temporal.

SPE Audit reúne la petición que se conserva, la regla aplicada y el desacuerdo observable en una misma unidad que otra persona puede revisar. El catálogo de **116 referencias** se publica aparte. Las entradas administrativas y comerciales conservan su identificador y su función documental. En 1.1 se examinaron de nuevo 58 referencias inicialmente mal fechadas. Se fechó un objeto identificable en 53 casos. En cinco persistió la incertidumbre. Una captura o una actualización fecha una versión concreta.

Uso en una reclamación. El expediente permite contrastar la salida impugnada, el corte y las fichas guardadas. El revisor debe distinguir fecha del suceso, publicación, entrada en vigor y captura. Los artículos 12 y 14 del Reglamento de IA de la UE sitúan las preguntas sobre registro y supervisión humana en sistemas de alto riesgo. La conformidad de un despliegue exige una evaluación propia.

Catálogo de antecedentes · Registro de fechas

Revisión independiente y reproducción

34 sucesos candidatos. El registro para el siguiente ensayo contiene 28 hechos realizados, cinco cancelaciones documentadas y un aplazamiento. Las páginas fuente están en inglés (19), español (10), francés (4) y alemán (1). Es una selección dirigida. Varios sucesos pueden compartir una misma causa. Las fuentes describen los hechos y las fechas. Todavía faltan las preguntas, las referencias de evaluación y las predicciones de esa cohorte.

Una ficha canaria ilustra la decisión que habrá que tomar al redactarlas. La Fundación César Manrique inauguró su sede en Tahíche el 27 de marzo de 1992 y abrió al público el 1 de abril. Para un corte del 30 de marzo, «se inauguró» y «abrió al público» merecen decisiones distintas. Confundir los dos hitos introduciría el error en la referencia antes de evaluar a ningún modelo.

Piloto humano abierto. Los participantes pueden utilizar el kit de 16 casos para recoger dos decisiones independientes por caso y recurrir a un tercer revisor en los desacuerdos. Se conserva la exposición previa a las respuestas. A fecha de esta edición, el proyecto ha medido cero decisiones y cero minutos de revisión humana. Los datos publicados tampoco permiten estimar el acuerdo entre revisores ni atribuir un coste monetario a las remisiones.

Para reproducir las tablas. Desde la carpeta v12, ejecutar `python -B score.py` con Python 3.11 o posterior. El programa contrasta el plan, lee los recibos archivados y recompone las decisiones y los recuentos. El repositorio conserva además el código y las instrucciones para una nueva inferencia local con los pesos publicados. Una repetición externa puede informar de la versión, la huella del plan y cualquier diferencia por identificador.

Kit de reproducción · Registro de 34 sucesos · Piloto abierto

Cita. Rubén Abella (2026). *Auditoría temporal de las respuestas de IA*. SPE Audit, edición 1.2. DOI 10.5281/zenodo.24447732. Texto y figuras originales bajo CC BY 4.0. El código se distribuye bajo la licencia BSL 1.1 del repositorio. La marca y el material de terceros conservan sus condiciones propias.