COLECCIÓN

MITOS Y REALIDADES DE LA IA

Entre la promesa y la realidad

POR QUÉ LA IA PUEDE PRODUCIR INFORMACIÓN FALSA CON APARIENCIA CONVINCENTE

Número 13



MRCI

anejo de ecursos y ontroles nteligentes

DR. JOSÉ ANTONIO PÉREZ RAMOS

POR QUÉ LA IA PUEDE PRODUCIR INFORMACIÓN FALSA CON APARIENCIA CONVINCENTE

SOBRE EL AUTOR

Doctor en Ciencias de lo Fiscal por el Instituto de Especialización para Ejecutivos. Maestro en Fiscal por la Universidad Autónoma "Benito Juárez" de Oaxaca. Maestro en Estudios Avanzados en Derechos Humanos por la Global Open University. Licenciado en Contaduría Pública por la Universidad Autónoma "Benito Juárez" de Oaxaca y Licenciado en Filosofía por la Universidad Nacionalista de México.

Socio fundador y director general de la Firma Manejo de Recursos y Controles Inteligentes (Firma MRCI). Fiscalista del año 2009 por la revista *Defensa Fiscal*. Doctor honoris causa por la Asamblea General del Claustro Doctoral Iberoamericano (2016). Doctor honoris causa por el Instituto Mexicano de Líderes de Excelencia (2017). Doctor honoris causa por 1Million Startups, Latinomics, Leaderships Forum y la Fundación Humanist World (2018). Doctor honoris causa por el Claustro de Doctores de la AISOL (American International School of Law) (2019).

Autor de diversas obras, entre las que destacan *Ius mutantur o del derecho que cambia* (Tirant Lo Blanch, 2025), *Futurología* (MRCI, 2025) y *Hasta que la muerte nos separe* (MRCI, 2025). Ha participado como coordinador en *Cuestiones tributarias. Problemas y controversias en el México actual* (Tirant Lo Blanch, 2023), *El costo de la justicia* (Apex Iuris, 2019), *Remuneraciones estratégicas inteligentes* (MRCI, 2015) y *Retos de la justicia tributaria* (Tirant Lo Blanch-AMDF, 2025).

Titular de *Pie de Página*, programa para la divulgación de los derechos humanos en México, a través de la plataforma Legal Cloud. Expositor en foros nacionales e internacionales. Colaborador en diversas revistas jurídicas, entre las que destacan *Defensa Fiscal* y *Legal Cloud*.

POR QUÉ LA IA PUEDE PRODUCIR INFORMACIÓN FALSA CON APARIENCIA CONVINCENTE

DR. JOSÉ ANTONIO PÉREZ RAMOS

PRIMERA EDICIÓN, ATEMPORAL

Derechos reservados, propiedad de
Dr. José Antonio Pérez Ramos

Comentarios y opiniones: investigacion@mrci.com.mx

Título original: Por qué la IA puede producir información falsa con apariencia convincente.

Colección mitos y realidades de la inteligencia artificial.

Entre la promesa y la realidad.

Autor: Dr. José Antonio Pérez Ramos.

Queda prohibida la reproducción total y parcial de esta obra denominada; POR QUÉ LA IA PUEDE PRODUCIR INFORMACIÓN FALSA CON APARIENCIA CONVINCENTE, por cualquier medio, sin autorización escrita del autor.

PRINTED IN MEXICO
IMPRESO EN MÉXICO

ÍNDICE

LA PREGUNTA QUE UNA RESPUESTA ELEGANTE NO RESUELVE ..... - 7 -

MÉTODO Y LÍMITES DE LA INVESTIGACIÓN..... - 9 -

QUÉ SIGNIFICA AFIRMAR ALGO FALSO ..... - 10 -

CÓMO SE CONSERVA LA FLUIDEZ CUANDO FALLA EL CONTENIDO ..... - 15 -

EXPLICACIONES QUE NO DEBEN CONFUNDIRSE CON UNA CAUSA ÚNICA - 18 -

PREGUNTAS QUE PARECEN CONTESTABLES Y NO LO SON..... - 21 -

UN TEXTO PUEDE MEZCLAR ACIERTOS Y ERRORES ..... - 23 -

RECUPERAR DOCUMENTOS AYUDA CUANDO SE CONSERVA LA RELACIÓN CON ELLOS ..... - 26 -

REPETICIÓN Y SEGURIDAD VERBAL NO SON PRUEBAS SUFICIENTES..... - 28 -

TRES MITOS QUE DIFICULTAN EL DIAGNÓSTICO..... - 31 -

CASO HIPOTÉTICO DE UNA INSPECCIÓN CONVERTIDA EN REPARACIÓN .. - 32 -

CASO HIPOTÉTICO DE DOS PERSONAS CON EL MISMO NOMBRE ..... - 34 -

CASO HIPOTÉTICO DE UNA PREGUNTA SIN DECISIÓN DOCUMENTADA..... - 36 -

UN PROTOCOLO DE COMPROBACIÓN PROPORCIONAL..... - 38 -

CASO HIPOTÉTICO DE EVALUACIÓN CON ABSTENCIONES ..... - 41 -

QUÉ PERMITEN AFIRMAR LAS PRUEBAS Y QUÉ PERMANECE ABIERTO..... - 45 -

CONCLUSIÓN..... - 48 -

REFERENCIAS..... - 49 -

La pregunta que una respuesta elegante no resuelve

El problema aparece cuando una persona recibe un informe legible y debe decidir si puede actuar con base en él; la buena sintaxis facilita entender lo que el sistema afirma, pero no muestra cómo se establecieron los hechos, si se preservaron las condiciones de una fuente o si un detalle fue completado sin evidencia suficiente.

Imaginemos una respuesta que describe una inspección de maquinaria con fechas, responsables y conclusiones ordenadas; aunque todos esos elementos resulten verosímiles, todavía habría que comprobar si la inspección ocurrió, si el responsable corresponde al equipo mencionado y si el documento realmente autorizó reanudar la operación, pues cada relación introduce una afirmación distinta.

La pregunta central es por qué un sistema capaz de producir lenguaje coherente puede fallar en esas relaciones sin presentar señales lingüísticas evidentes de error; responder exige separar lo que se observa en la salida, lo que se conoce del procedimiento técnico y lo que sólo puede proponerse como explicación de un caso particular.

La tesis de este artículo es que generar una continuación apropiada y sostener una afirmación mediante evidencia son tareas relacionadas, pero no equivalentes; un modelo puede mejorar en una sin resolver completamente la otra, por lo que la calidad de una respuesta factual debe establecerse mediante criterios adicionales a su forma verbal.

El alcance se concentra en sistemas que generan texto, especialmente modelos de lenguaje y aplicaciones que los combinan con documentos recuperados; no se presume que toda inteligencia artificial produzca los mismos errores, ni se trasladan automáticamente estas conclusiones a un clasificador de imágenes, un sistema experto de reglas o un controlador industrial.

La expresión «apariencia convincente» se emplea aquí para describir una salida fluida, pertinente en su tema y presentada con detalles que parecen verificables; no equivale a una medición experimental de persuasión en todas las personas, porque demostrar ese efecto exigiría estudiar lectores, condiciones de presentación y decisiones efectivamente tomadas.

Método y límites de la investigación

Se realizó una revisión documental narrativa de ocho investigaciones primarias, con fecha de corte del 3 de octubre de 2026; la selección cubre distinciones conceptuales, pruebas de comportamiento y propuestas de mitigación, sin constituir una búsqueda sistemática exhaustiva de toda la literatura ni un metaanálisis de tasas comparables.

Las publicaciones sobre resumen y preguntas sin respuesta permiten observar fallos en tareas acotadas, mientras las de generación extensa muestran dificultades al reunir múltiples afirmaciones; se incluye una prepublicación teórica de 2025 en su primera versión para analizar una explicación sobre incentivos, conservando la diferencia entre un argumento formal y la identificación empírica de una causa dentro de un producto concreto.

Cada resultado se interpreta en relación con la tarea, los modelos y el criterio de evaluación empleados por sus autores; los nombres de sistemas que aparecen en los estudios identifican antecedentes históricos y no representan una clasificación de los productos disponibles

en la fecha de corte, cuyas configuraciones pueden haber cambiado.

Los casos del artículo son hipotéticos y sus personas, documentos y cantidades fueron construidos para explicar decisiones de verificación; no describen experimentos nuevos con modelos, incidentes reales ni resultados obtenidos por una empresa, además los cálculos ilustrativos dependen exclusivamente de los supuestos declarados en cada caso.

El término «evidencia» comprende aquí pasajes que sostienen una proposición y observaciones documentadas que permiten evaluarla; una referencia auténtica puede resultar irrelevante para una afirmación, por lo que verificar la existencia de una fuente y verificar lo que esa fuente acredita son operaciones distintas.

Qué significa afirmar algo falso

Una afirmación factual pretende describir un estado de cosas susceptible de contraste, como una fecha, una cantidad, una identidad o la relación entre un hecho y otro; su evaluación requiere conservar a qué persona, momento



y situación se refiere, porque retirar esas condiciones puede transformar una frase correcta en una conclusión equivocada.

Llamaremos exactitud factual a la correspondencia de la afirmación con los hechos que se intenta describir, atendiendo a la evidencia pertinente; emplearemos fidelidad para indicar que una respuesta conserva lo que sus materiales de referencia permiten sostener, sin añadirles relaciones que no contienen ni deformar las que sí contienen.

Maynez et al. (2020, pp. 1907–1909) distinguen en el resumen automático el contenido no respaldado por el documento y la posibilidad de que una adición sea, pese a ello, factualmente correcta; también separan alteraciones de información presente de incorporaciones que no pueden inferirse directamente del texto de entrada.

Esa distinción evita llamar falso a todo lo que no aparece en un archivo, aunque también impide atribuir a ese archivo información que procede de otra parte; si el encargo consiste en resumir exclusivamente un acta, una adición verdadera pero ajena al acta puede incumplirlo,



mientras una síntesis contextualizada podría admitirla si identifica su procedencia.

La relación inversa también importa: una reproducción fiel de un documento equivocado conserva lo que el documento dice, pero no demuestra que su contenido describa correctamente el mundo; cuando se solicita informar qué afirma una fuente, la atribución puede ser correcta aunque resulte necesario advertir que existe evidencia contraria.

La tabla 1 organiza las decisiones que surgen de estas distinciones mediante ejemplos originales; sus categorías son instrumentos de análisis y pueden superponerse en una misma respuesta, de modo que no pretenden constituir una clasificación exhaustiva de todos los errores posibles.

Tabla 1

Distinciones necesarias antes de corregir una respuesta

Situación	Qué debe comprobarse	Respuesta proporcionada
Contradicción documental	Si el texto fuente afirma algo incompatible con la salida	Corregir la proposición y revisar sus consecuencias
Adición sin respaldo	Si existe una fuente independiente que sostenga el detalle	Identificar la nueva fuente o retirar la afirmación
Fuente posiblemente errónea	Si el documento es fielmente resumido pero está desmentido	Conservar la atribución y explicar el conflicto

Situación	Qué debe comprobarse	Respuesta proporcionada
Ambigüedad de identidad	Si todos los datos corresponden a la misma entidad	Separar entidades o pedir el identificador faltante
Pregunta no contestable	Si el material permite obtener lo solicitado	Delimitar lo conocido y abstenerse del detalle ausente

Nota. Elaboración propia; la distinción entre fidelidad y exactitud se apoya en Maynez et al. (2020), mientras los controles de identidad y contestabilidad se relacionan con Chiang y Lee (2024) y Rajpurkar et al. (2018), respectivamente.

«Sin respaldo» debe conservarse como una categoría independiente cuando el material disponible no permite



confirmar ni refutar una proposición; convertir automáticamente esa ausencia en falsedad confunde el estado de nuestra comprobación con el estado del mundo, y puede producir una segunda afirmación injustificada al intentar corregir la primera.

Tampoco toda discrepancia es un error factual: una recomendación puede depender de criterios de preferencia, una estimación puede expresar incertidumbre y una hipótesis puede formular una explicación todavía no probada; el problema aparece cuando esas categorías se presentan como hechos establecidos o cuando se omiten las condiciones que permiten interpretarlas.

Cómo se conserva la fluidez cuando falla el contenido

Un modelo de lenguaje ajusta relaciones estadísticas entre representaciones de texto y utiliza lo aprendido para producir una respuesta condicionada por la entrada; que una secuencia resulte probable bajo ese procedimiento no significa que cada proposición de la secuencia haya sido contrastada con un registro independiente del mundo.

La competencia lingüística permite conservar patrones de un informe, enlazar conceptos y completar estructuras habituales; por ello es posible que un detalle incorrecto ocupe un lugar gramatical adecuado, se relacione con el tema general y parezca compatible con el resto del párrafo, aunque no cuente con el respaldo que el lector supone.

En su estudio de resumen extremo, Maynez et al. (2020, pp. 1906–1908) documentaron salidas fluidas y relacionadas con el tema que alteraban entidades o relaciones del documento; además, explican que optimizar la probabilidad de un resumen de referencia no equivale necesariamente a recompensar su fidelidad, especialmente cuando los propios pares de entrenamiento incluyen información no compartida.

El hallazgo apoya una afirmación limitada pero importante: la incorrección puede coexistir con una forma verbal competente en sistemas observados; no demuestra que todas las respuestas fluidas sean sospechosas en igual grado, ni que una arquitectura particular esté condenada a producir siempre un porcentaje fijo de información falsa.

La precisión aparente también puede surgir de la forma del encargo, como cuando se exige una fecha exacta, una explicación causal completa o un listado con un número obligatorio de elementos; si el sistema completa esa estructura sin reconocer sus límites, el formato terminado puede esconder que algunos espacios no tenían información suficiente para llenarse.

Conviene distinguir el diagnóstico observable de la explicación interna: podemos verificar que el sistema inventó una fecha, pero concluir que «recordó mal», «quiso complacer» o «ocultó su ignorancia» requiere evidencia adicional; esas expresiones humanas suelen condensar hipótesis diferentes y no identifican por sí mismas qué componente falló.

Una respuesta razonable ante esa diferencia consiste en pedir trazabilidad de las afirmaciones relevantes, no una narración introspectiva del modelo; señalar el pasaje que respalda una fecha o ejecutar un cálculo independiente aporta una prueba revisable, mientras una explicación verbal de por qué el sistema se considera seguro sigue siendo otra salida que necesita evaluación.

Explicaciones que no deben confundirse con una causa única

Las causas de un error pueden situarse en distintos puntos del proceso: datos inadecuados, aprendizaje insuficiente para una relación, interpretación errónea de la pregunta, combinación defectuosa de documentos o criterios que favorecen responder aun sin respaldo; enumerar esas posibilidades ayuda a investigar, pero no permite atribuir cualquiera de ellas a un caso sin realizar comprobaciones.

Lin et al. (2022, pp. 3214–3217) diseñaron TruthfulQA para examinar respuestas que reproducen falsedades y concepciones equivocadas presentes en el discurso humano; su trabajo plantea la imitación de esas regularidades como una explicación relevante y contrasta el desempeño en preguntas adversariales con otros tipos de preguntas.

La contribución no consiste en probar que todo error fue copiado literalmente de una fuente de entrenamiento, sino en mostrar que imitar mejor determinadas distribuciones de texto no garantiza responder con mayor veracidad; sus resultados pertenecen a los modelos y condiciones

estudiados, por lo que no justifican la regla universal de que un modelo mayor será menos exacto.

Existe otra posibilidad: aunque los ejemplos observados durante el aprendizaje sean correctos, la información disponible puede ser insuficiente para determinar un dato nuevo; conocer muchas fechas de apertura de centros culturales no permite deducir la fecha exacta de un centro particular que nunca se observó, porque una regularidad de formato no identifica el valor ausente.

Kalai et al. (2025, secciones 1 y 3) formalizan relaciones entre errores de generación y dificultades para distinguir salidas válidas de inválidas; su análisis incluye situaciones con datos de entrenamiento correctos y hechos arbitrarios, bajo supuestos explícitos, de modo que la limpieza de datos no aparece como garantía suficiente para evitar cualquier error de generalización.

El alcance de ese argumento debe mantenerse acotado: una demostración bajo determinadas condiciones no constituye una medición del mecanismo interno de cada asistente, ni implica que cualquier tarea sea igualmente difícil; tareas restringidas pueden resolverse mediante



reglas, consultas exactas o procedimientos que permitan rechazar resultados no comprobados.

En la misma prepublicación, Kalai et al. (2025, secciones 1.2 y 4) proponen que las evaluaciones que premian aciertos y asignan el mismo resultado a abstenciones y errores favorecen estrategias de adivinación; ésta es una explicación sobre incentivos de desarrollo y selección, no una atribución de deseos conscientes al sistema evaluado.

Podemos ilustrar el incentivo sin identificarlo con un experimento: si una prueba otorga un punto por acertar y cero tanto por equivocarse como por abstenerse, intentar una respuesta con alguna probabilidad de acierto tiene mayor puntuación esperada que callar; en una aplicación donde una respuesta falsa genera un costo, esa preferencia puede invertirse.

La conclusión práctica es investigar qué se optimiza y qué se mide antes de atribuir el problema a «la IA» como un conjunto uniforme; un cambio de interfaz, una instrucción de responder siempre o una selección de documentos diferente puede modificar los errores de una aplicación sin alterar los parámetros del modelo que la sustenta.

Preguntas que parecen contestables y no lo son

Una pregunta bien escrita puede contener una premisa no acreditada, pedir una relación inexistente o exigir información que no se encuentra en el material disponible; el sistema necesita evaluar esa condición antes de elegir una respuesta, porque encontrar una frase relacionada con el tema no basta para demostrar que responde a lo solicitado.

Rajpurkar et al. (2018, pp. 784–786) construyeron preguntas sin respuesta que resultaban pertinentes al documento y contenían candidatos plausibles; el diseño obligaba a los sistemas de lectura a distinguir entre localizar un fragmento del tipo esperado y establecer que el texto realmente sustentaba la respuesta.

Su contribución permite separar dos capacidades evaluables: responder cuando la información existe y reconocer cuándo no está sustentada; el estudio pertenece a sistemas de comprensión extractiva y no proporciona una tasa universal de error para asistentes actuales, aunque su distinción sigue siendo útil para diseñar pruebas de aplicaciones documentales.

En un acta que dice «se discutió cambiar de proveedor», preguntar «¿qué proveedor fue contratado?» introduce una decisión que todavía no consta; elegir el nombre más cercano puede producir una respuesta breve y convincente, pero el error principal estaría en aceptar la premisa de contratación sin verificarla.

La abstención adecuada puede ser localizada: es posible informar que hubo discusión, identificar los proveedores considerados y aclarar que no se documentó una adjudicación; responder parcialmente con límites explícitos conserva utilidad, mientras una negativa indiscriminada desperdiciaría información que sí está respaldada.

También debe distinguirse «el documento no lo indica» de «no ocurrió»: la primera frase describe una carencia comprobable del material examinado y la segunda introduce una conclusión sobre los hechos; si el archivo no es exhaustivo, no puede utilizarse su silencio como prueba suficiente de inexistencia.

Un procedimiento de evaluación que sólo incluya preguntas con respuesta conocida deja sin examinar este

límite; conviene incorporar solicitudes con información ausente, nombres parecidos y premisas equivocadas, siempre con una respuesta de referencia que explique por qué corresponde contestar, aclarar o abstenerse.

Un texto puede mezclar aciertos y errores

Una respuesta extensa rara vez contiene una sola proposición; una oración puede afirmar simultáneamente el nombre de una persona, su cargo, el año de un nombramiento y la institución correspondiente, por lo que aceptar o rechazar todo el párrafo con una etiqueta única puede ocultar dónde está el problema y qué parte debe corregirse.

Min et al. (2023, pp. 12076–12079) propusieron FActScore para descomponer textos en afirmaciones atómicas y evaluar qué proporción está respaldada por una fuente de conocimiento; el trabajo examinó biografías y mostró la utilidad de distinguir piezas sustentadas y no sustentadas dentro de una misma generación.

Una afirmación atómica es, para ese propósito, una unidad que comunica un elemento de información verificable;



separar «Elena dirigió el museo desde 2018» permite revisar la identidad de Elena, su relación con el museo y la fecha de inicio, aunque debe conservarse la vinculación entre esos elementos para evitar que cada uno termine describiendo personas distintas.

La métrica de Min et al. (2023, sección 3.1) mide precisión respecto de una fuente elegida y no cobertura de toda la información que habría sido útil comunicar; además, sus supuestos sobre peso equivalente de los hechos y consistencia de la fuente advierten por qué una puntuación no puede trasladarse sin más a cualquier decisión profesional.

De esa limitación se deriva una precaución editorial: una respuesta que omite el dato difícil puede obtener una proporción alta de afirmaciones respaldadas y aun así no resolver la necesidad del usuario; la evaluación debe preguntar tanto si lo escrito es correcto como si falta un elemento indispensable para actuar.

La gravedad tampoco se distribuye de forma uniforme: equivocarse en una fecha periférica y confundir si una prueba confirmó o descartó una condición no tienen

necesariamente la misma consecuencia; contar aciertos ayuda a localizar patrones, pero decidir si una respuesta es utilizable exige examinar qué depende de cada afirmación.

Chiang y Lee (2024, pp. 2734–2736) muestran una dificultad adicional mediante biografías de nombres ambiguos: afirmaciones verificables por separado pueden reunirse en un párrafo que induce a atribuirlas a una misma persona cuando proceden de entidades distintas; su evaluación incorpora la desambiguación para examinar esa combinación.

Este resultado justifica revisar la coherencia referencial del conjunto además de las piezas aisladas; una fecha de nacimiento, una profesión y una obra pueden existir en fuentes auténticas, pero su unión carecer de respaldo si describen a personas diferentes, de modo que la autenticidad documental no garantiza la corrección de la relación construida.

La revisión tiene, por tanto, dos escalas complementarias: localizar el apoyo de cada afirmación y reconstruir lo que el lector entenderá del conjunto; la segunda escala permite

detectar cambios de sujeto, cronologías incompatibles, generalizaciones indebidas y conclusiones que ninguna de las fuentes formuló o permitió inferir.

Recuperar documentos ayuda cuando se conserva la relación con ellos

La recuperación aumentada por generación combina un componente que localiza materiales con otro que produce una respuesta utilizando esos materiales; Lewis et al. (2020, secciones 2 y 4) estudiaron una arquitectura de este tipo y documentaron mejoras frente a una línea de base sin recuperación en tareas y comparaciones concretas.

La mejora observada respalda el valor de aportar evidencia externa, pero no autoriza concluir que todo sistema conectado a un buscador responde correctamente; la intervención del generador sigue requiriendo controles sobre qué fragmentos utiliza, qué relaciones conserva y qué agrega al redactar.

En términos de análisis del proceso, conviene separar la recuperación de una fuente pertinente, la lectura de su pasaje relevante y el apoyo efectivo de la conclusión; un

documento puede superar la primera comprobación y fallar en la tercera, como cuando aborda el mismo asunto pero sólo describe una propuesta y la respuesta la presenta como una decisión ejecutada.

El estudio de Chiang y Lee (2024, secciones 2 y 3) resulta especialmente pertinente porque examina generaciones apoyadas en pasajes recuperados y encuentra combinaciones problemáticas entre entidades; disponer de documentos relacionados con el nombre buscado no resuelve automáticamente cuál de sus datos pertenece al referente solicitado.

Un sistema también puede disponer de una fuente correcta y responder con un detalle incompatible con ella; para verificar ese caso hace falta comparar el texto producido con el pasaje, pues la presencia de un enlace, un título reconocible o un indicador visual de búsqueda sólo acredita una parte del recorrido.

La solución no consiste necesariamente en exigir más documentos, porque añadir materiales puede introducir versiones antiguas, homónimos o puntos de vista incompatibles; la cantidad de evidencia importa cuando



aumenta el respaldo pertinente, mientras acumular coincidencias temáticas puede elevar el trabajo de desambiguación sin fortalecer la conclusión.

Una respuesta verificable debería permitir distinguir el hecho directamente respaldado, la inferencia que conecta varios hechos y el elemento pendiente; esa separación no requiere exponer todos los detalles técnicos de la aplicación, pero sí evita presentar como una transcripción documental aquello que en realidad fue una elaboración adicional.

Repetición y seguridad verbal no son pruebas suficientes

Si un sistema responde varias veces lo mismo, sabemos que la respuesta se mantuvo bajo las condiciones probadas; no sabemos todavía si describe los hechos, porque un procedimiento estable puede repetir un error y una modificación de redacción puede conservar un contenido correcto.

La temperatura y otros parámetros de generación afectan la selección de salidas, pero reducir la variación no



introduce por sí mismo una comprobación factual; si la continuación favorecida contiene un dato incorrecto, una generación más repetible puede seguir produciéndolo, sin que de esa observación se derive una ley universal sobre todas las configuraciones.

Farquhar et al. (2024, pp. 625–629) proponen detectar un subconjunto de errores mediante entropía semántica, agrupando respuestas por su significado en lugar de tratar cada redacción como diferente; su método busca identificar variaciones de contenido compatibles con respuestas arbitrarias y señala expresamente que no garantiza exactitud cuando el sistema se equivoca de manera sistemática.

La aportación permite comprender por qué la consistencia aporta información sin resolver por completo la verificación: discrepancias entre respuestas pueden advertir un problema, pero la ausencia de discrepancias no descarta un error compartido; tampoco toda variación semántica es indebida si la pregunta admite varias respuestas válidas o requiere una aclaración.



Preguntar «¿estás seguro?» puede cambiar el tono, la extensión o el contenido de una respuesta, pero no debe contarse como prueba independiente sólo por aparecer en un segundo turno; lo que debe examinarse es si la revisión incorporó una fuente, una operación o un criterio de contraste que permita decidir entre las alternativas.

Pedir a otro modelo una opinión puede ser útil para detectar contradicciones, aunque el acuerdo entre ambos tampoco acredita el hecho; si comparten una confusión de identidad o dependen del mismo documento erróneo, sus respuestas coincidentes no equivalen a dos observaciones independientes del acontecimiento investigado.

La seguridad verbal merece el mismo tratamiento: expresiones categóricas y porcentajes de confianza sólo resultan útiles cuando se conoce cómo se obtuvieron y qué relación mantienen con errores observados; sin esa comprobación, deben interpretarse como parte de la respuesta, no como una garantía externa que la valida.



Tres mitos que dificultan el diagnóstico

El primer mito consiste en llamar mentira a cualquier información falsa generada; la falsedad describe una relación entre una afirmación y los hechos, mientras atribuir engaño deliberado añade una explicación sobre intención que no se desprende únicamente del contenido erróneo, aunque el efecto práctico sobre una persona pueda ser perjudicial.

Evitar esa atribución automática no minimiza el riesgo ni elimina responsabilidades humanas por el uso de un sistema; permite formular una investigación más precisa sobre la fuente, el entrenamiento, la interfaz y los controles, además de separar un fallo espontáneo de una aplicación diseñada por alguien para difundir información engañosa.

El segundo mito sostiene que una respuesta larga y detallada debe contener más conocimiento comprobado; la extensión añade oportunidades de explicar y contextualizar, pero también añade afirmaciones que necesitan respaldo, por lo que un informe más breve puede resultar preferible si conserva lo relevante y elimina detalles que nadie puede verificar.

El tercer mito afirma que una advertencia genérica al final vuelve segura cualquier respuesta; informar que un sistema puede equivocarse no identifica cuáles de sus afirmaciones son inciertas ni resuelve qué debe comprobar el lector, de modo que la advertencia sólo ayuda cuando acompaña límites concretos y un procedimiento utilizable.

Estas objeciones no proponen desconfiar de manera uniforme, porque una sospecha indiferenciada consume tiempo y puede impedir tareas de bajo riesgo; proponen que la confianza dependa del tipo de afirmación, del uso previsto y de las comprobaciones disponibles, con especial cuidado cuando una presentación convincente oculta una dependencia crítica.

Caso hipotético de una inspección convertida en reparación

Una organización ficticia entrega al asistente un parte que dice: «El equipo B fue inspeccionado el martes; se observó desgaste y se recomendó sustituir una pieza antes de reiniciar su uso»; al solicitar un resumen ejecutivo, recibe una frase según la cual el equipo fue reparado el martes y quedó disponible para operar.

El resumen conserva el equipo, el día y el tema de mantenimiento, pero cambia dos relaciones decisivas: transforma una recomendación en una acción concluida y atribuye una autorización de uso que el parte no contiene; la fluidez de la frase no permite detectar ese cambio sin comparar los verbos y sus condiciones.

La primera comprobación compara tres afirmaciones del parte y del resumen: hubo inspección, hubo reparación y se autorizó el uso; sólo la primera aparece respaldada por el material descrito, mientras la segunda y la tercera requieren registros adicionales que no deben darse por existentes.

La corrección proporcionada sería informar la inspección, conservar la observación de desgaste y señalar que el documento recomienda una sustitución previa al reinicio; si la decisión práctica depende de saber si la reparación se ejecutó, debe solicitarse el registro correspondiente en lugar de inferir su realización a partir de la recomendación.

Este ejemplo no demuestra cuál mecanismo interno produjo el error, porque podría intervenir una lectura deficiente, una simplificación excesiva o una instrucción

que favoreció un resumen concluyente; sí permite identificar qué perdió fidelidad y qué evidencia adicional sería necesaria para emitir una conclusión operativa.

La revisión también debe propagarse a las consecuencias: retirar la frase «equipo disponible» no basta si un tablero, una orden de trabajo o un mensaje posterior ya la utilizaron; dentro del alcance autorizado, conviene localizar esas dependencias para evitar que una corrección textual deje activa la decisión equivocada.

Caso hipotético de dos personas con el mismo nombre

Un equipo editorial prepara una nota sobre una investigadora ficticia llamada Clara Montes y recibe dos documentos: uno describe a Clara Montes, investigadora de materiales desde 2016, y otro a Clara Montes, directora de un archivo histórico desde 2021; ninguno indica que se trate de la misma persona.

El asistente genera una biografía que atribuye a una sola Clara Montes ambas actividades y explica una transición profesional entre ellas; esa transición hace que el texto



resulte coherente, pero constituye una relación añadida, pues los documentos sólo proporcionan coincidencia nominal y trayectorias separadas.

Consultar de nuevo ambos archivos podría confirmar cada cargo sin resolver la identidad, por lo que una revisión centrada exclusivamente en palabras coincidentes aprobaría hechos aislados y conservaría la unión incorrecta; el control necesario consiste en encontrar un identificador que vincule o separe a las personas, no en acumular más menciones del nombre.

Una respuesta prudente indicaría que hay dos registros que pueden corresponder a personas distintas y limitaría la nota al referente identificado por el solicitante; si no existe información suficiente para distinguirlo, preguntaría por institución, área o documento de referencia, evitando divulgar una trayectoria compuesta como si estuviera acreditada.

La enseñanza coincide con el problema de composición examinado por Chiang y Lee (2024), pero los nombres y circunstancias de este caso fueron creados para el artículo; su función es mostrar cómo una narración puede

incorporar hechos individualmente respaldados y producir una conclusión biográfica que no lo está.

En trabajos extensos, conviene mantener el identificador de cada entidad durante toda la revisión y comprobarlo cuando se integren fuentes; cambiar de documento no debe cambiar inadvertidamente de persona, de producto o de versión de una organización, aunque las denominaciones parezcan equivalentes.

Caso hipotético de una pregunta sin decisión documentada

Un comité ficticio registra que dos propuestas de actividades serán evaluadas la semana siguiente y que cada responsable enviará información adicional antes del viernes; el usuario pide al asistente la fecha definitiva de inicio de la propuesta seleccionada, aun cuando el acta no incluye selección ni calendario aprobado.

El sistema podría elegir ese viernes y presentarlo como inicio definitivo, pero eso confundiría el plazo para recibir información con el inicio de la actividad; incluso si la fecha



se copia exactamente, la respuesta atribuiría a ese dato una función que el documento no le asigna.

La verificación debe comenzar por las premisas de la solicitud: comprobar si hubo selección y si se aprobó una fecha; cuando ambas condiciones faltan, la respuesta útil consiste en informar el plazo efectivamente registrado y señalar qué decisión permanece pendiente, sin convertir una expectativa en un hecho.

Este caso muestra por qué la extracción literal tampoco garantiza una respuesta correcta: una cadena de palabras puede existir en el documento y ser inadecuada como contestación a una pregunta; la comprobación requiere conservar la relación entre el dato extraído, su sujeto y el acontecimiento que describe.

Una prueba de calidad podría presentar variantes del mismo caso con una selección explícita, con una fecha tentativa y con información ausente; el objetivo sería observar si el sistema distingue las condiciones, en vez de premiar que siempre produzca una fecha con el formato solicitado.

Un protocolo de comprobación proporcional

La tabla 2 presenta una propuesta editorial de comprobación, no un estándar validado experimentalmente ni una obligación idéntica para todas las tareas; su finalidad es elegir un control que responda al riesgo observado, procurando que la revisión produzca una decisión y no sólo una lista de enlaces.

Tabla 2. *Controles propuestos según la consecuencia del uso*

Uso previsto	Comprobación mínima razonada	Condición para detenerse
Exploración de ideas	Separar propuestas de hechos y revisar los datos que se reutilizarán	Una afirmación pasa a sostener una decisión concreta
Comunicación informativa	Cotejar nombres, fechas, cifras y relaciones relevantes	No se encuentra respaldo para una afirmación central

Uso previsto	Comprobación mínima razonada	Condición para detenerse
Documento profesional	Revisar afirmaciones decisivas, versiones y dependencias	Existe contradicción material o fuente insuficiente
Decisión difícil de revertir	Obtener evidencia competente y revisión especializada apropiada	La consecuencia depende de un dato todavía no confirmado

Nota. Elaboración propia; los niveles describen usos hipotéticos y no sustituyen los requisitos aplicables a una actividad especializada.

El primer paso es definir qué se hará con la respuesta, porque un texto destinado a explorar posibilidades y uno utilizado para autorizar una operación requieren controles distintos; esa definición permite identificar las afirmaciones cuya corrección cambiaría la decisión y evitar dedicar el mismo esfuerzo a cada detalle ornamental.

El segundo paso consiste en separar hechos, inferencias, propuestas y datos pendientes; una inferencia puede ser razonable, pero debe conservar las premisas que la sostienen y el alcance que permiten, mientras una propuesta no debe describirse como decisión ya adoptada ni un dato ausente convertirse en una cifra estimada sin aviso.

El tercer paso es localizar la evidencia de cada afirmación decisiva y leer el pasaje con suficiente contexto; conviene registrar sujeto, fecha, condición y versión para comprobar que la fuente habla del mismo objeto, además de contrastar si la paráfrasis conserva negaciones, límites y excepciones.

El cuarto paso examina la composición: las piezas deben referirse a entidades compatibles, mantener el orden temporal y sostener la conclusión conjunta; una revisión puede aprobar cada dato por separado y aun así rechazar el argumento si sus conexiones fueron añadidas sin justificación.

El quinto paso utiliza comprobaciones independientes cuando la naturaleza del dato lo permite, como recalcular

una operación o comparar un total con sus componentes; la independencia se refiere al método de contraste y a la evidencia utilizada, no simplemente a que el mismo sistema redacte una segunda explicación.

El sexto paso resuelve el estado de la respuesta: conservar lo respaldado, corregir lo contradicho, identificar lo incierto y abstenerse del detalle que no puede establecerse; si el uso requiere ese detalle, la tarea permanece pendiente en lugar de cerrarse mediante una formulación más segura.

Por último, es necesario revisar la versión que realmente se utilizará, ya que un dato corregido puede volver a alterarse durante un resumen, una traducción o una integración posterior; el control pierde utilidad si sólo se aplica a un borrador y la salida final contiene relaciones nuevas que nadie examinó.

Caso hipotético de evaluación con abstenciones

Supongamos una evaluación ilustrativa de cien preguntas con respuestas de referencia revisadas y dos políticas de uso del mismo tipo de asistente; la política A responde



todas y obtiene ochenta aciertos y veinte errores, mientras la política B responde ochenta, acierta setenta y seis, se equivoca en cuatro y se abstiene en las veinte restantes.

Las cantidades son supuestos construidos y no resultados de una prueba ejecutada; la comparación sirve para mostrar que una sola métrica puede ocultar decisiones importantes, además no permite inferir que cualquier mecanismo de abstención producirá esa mejora ni que las preguntas omitidas tengan la misma dificultad que las respondidas.

La tabla 3 expone los cálculos: la cobertura corresponde a preguntas respondidas entre preguntas totales, mientras la exactitud entre respuestas emitidas divide los aciertos por el número de respuestas; también se conserva el número de errores, porque una decisión práctica puede depender de cuántas respuestas equivocadas llegan al usuario.

Tabla 3. *Comparación aritmética de dos políticas hipotéticas*

Indicador	Política A	Política B
Preguntas totales	100	100

Indicador	Política A	Política B
Respuestas emitidas	100	80
Aciertos	80	76
Errores	20	4
Abstenciones	0	20
Cobertura	100 %	80 %
Exactitud entre respuestas emitidas	80 %	95 %
Aciertos respecto del total	80 %	76 %

Nota. Elaboración propia con datos hipotéticos; 76 dividido entre 80 equivale a 0.95 y 80 dividido entre 100 equivale a 0.80, sin que estas proporciones representen desempeño observado de un producto.

Si sólo se contaran aciertos sobre el total, A parecería preferible por cuatro respuestas correctas adicionales; si una respuesta falsa tuviera una consecuencia importante, B podría ser más adecuada porque expone al usuario a



dieciséis errores menos, a cambio de dejar veinte preguntas sin resolver inicialmente.

La elección depende del costo de las abstenciones y de las equivocaciones, además de la posibilidad de revisar los casos pendientes; una aplicación que puede enviar veinte preguntas a una persona competente enfrenta una decisión distinta de otra donde la abstención paraliza una operación, por lo que no existe un umbral universal deducible de esta tabla.

La comparación también exige examinar qué preguntas se omiten, pues una política que se abstiene de manera desproporcionada ante nombres infrecuentes o formulaciones de ciertos usuarios podría ofrecer un servicio desigual; identificar esa distribución requeriría datos adicionales y no puede resolverse a partir de los totales hipotéticos presentados.

Este ejemplo desarrolla una consecuencia práctica del problema de puntuación analizado por Kalai et al. (2025), aunque el criterio de selección aquí propuesto pertenece al artículo; una evaluación útil debe evitar que la necesidad

de obtener una cifra única borre diferencias entre responder, equivocarse y reconocer que falta evidencia.

Qué permiten afirmar las pruebas y qué permanece abierto

Los estudios consultados permiten documentar que algunas generaciones reúnen fluidez y errores, que un texto puede mezclar afirmaciones respaldadas y no respaldadas, y que la recuperación documental no elimina problemas de composición; permiten también identificar intervenciones y mediciones prometedoras dentro de los escenarios examinados.

Sin embargo, un conjunto de pruebas no equivale a una estimación de todas las respuestas que una persona recibirá durante su trabajo; cambian las preguntas, los idiomas, los documentos, las versiones de los modelos y las consecuencias de los errores, de modo que una aplicación requiere evaluación sobre tareas suficientemente parecidas a las que realizará.

El diseño de TruthfulQA hace visible esta limitación: sus preguntas buscan poner a prueba determinadas

falsedades y no representan una muestra aleatoria de cualquier conversación posible (Lin et al., 2022, sección 2); su valor está en examinar una vulnerabilidad definida, mientras su interpretación exige conservar ese objetivo.

La revisión humana también necesita criterios explícitos, porque dos revisores pueden discrepar acerca de una inferencia, la suficiencia de una fuente o la relevancia de una omisión; el desacuerdo debe registrarse y resolverse mediante evidencia o una delimitación de la pregunta, en vez de ocultarse dentro de un promedio de puntuaciones.

Una investigación causal más sólida sobre una aplicación concreta compararía condiciones controladas: la misma consulta con y sin documentos pertinentes, con identidades desambiguadas y con autorización expresa para abstenerse; observar diferencias ayudaría a aislar componentes, aunque seguiría siendo necesario controlar cambios simultáneos de modelo, instrucciones y evaluación.

Permanece abierta la cuestión de cómo reducir simultáneamente errores, abstenciones innecesarias y costos de revisión en dominios diversos; ningún resultado

aquí consultado permite afirmar que exista un procedimiento único capaz de resolver esa combinación para cualquier contenido, y los métodos de detección deben evaluarse también cuando señalan equivocadamente una respuesta correcta.

La objeción de que los seres humanos también se equivocan es válida, pero no decide qué sistema conviene utilizar; la comparación relevante debe considerar calidad, velocidad, capacidad de corregir, acceso a evidencia y consecuencias, evitando tanto exigir infalibilidad a una herramienta como tolerar errores evitables porque exista falibilidad humana.

Otra objeción señala que comprobar cada afirmación elimina el ahorro de tiempo; la respuesta propuesta es concentrar la revisión en los hechos que sostienen la decisión y restringir el alcance cuando verificarlos resulte demasiado costoso, sin prometer que toda tarea asistida conservará una ventaja económica después de incorporar controles adecuados.

Conclusión

La información falsa puede adquirir una apariencia convincente porque producir lenguaje fluido y establecer el respaldo de una afirmación no son la misma operación; la evidencia examinada muestra fallos de fidelidad, preguntas sin respuesta y combinaciones incorrectas de hechos, mientras las explicaciones sobre aprendizaje e incentivos requieren conservar sus supuestos y límites.

El diagnóstico útil comienza por precisar qué está mal: una contradicción, un detalle no acreditado, una identidad ambigua o una conclusión que excede sus premisas; esa precisión permite elegir controles que respondan al problema, en lugar de confiar exclusivamente en el tono, la extensión, la repetición o la presencia de enlaces.

La verificación proporcional combina lectura de fuentes, revisión de relaciones y comprobaciones independientes cuando son posibles; una respuesta puede ser parcialmente aprovechable si conserva lo respaldado y declara lo pendiente, pero no debe convertirse en fundamento de una decisión que depende justamente de la información ausente.

La meta razonable es conocer cuándo una salida resulta utilizable y cuándo debe corregirse, ampliarse o detenerse; evaluar exactitud junto con cobertura y abstención ayuda a sostener esa decisión sin atribuir intención humana a cada error ni prometer que una mejora técnica eliminará definitivamente las afirmaciones falsas.

Referencias

Chiang, C.-H., y Lee, H. (2024). Merging facts, crafting fallacies: Evaluating the contradictory nature of aggregated factual claims in long-form generations [Unir hechos y construir falacias: Evaluación de la naturaleza contradictoria de afirmaciones factuales agregadas en generaciones extensas]. En L.-W. Ku, A. Martins, y V. Srikumar (Eds.), *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 2734–2751). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.160

Farquhar, S., Kossen, J., Kuhn, L., y Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy [Detección de alucinaciones en grandes modelos de lenguaje mediante entropía semántica]. *Nature, 630*(8017), 625–630. https://doi.org/10.1038/s41586-024-07421-0

Kalai, A. T., Nachum, O., Vempala, S. S., y Zhang, E. (2025). *Why language models hallucinate* [Por qué los modelos de lenguaje alucinan] (Versión 1) [Prepublicación]. arXiv. https://doi.org/10.48550/arXiv.2509.04664

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., y Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks [Generación aumentada por recuperación para tareas de procesamiento del lenguaje natural intensivas en conocimiento]. En H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, y H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 9459–9474). Curran Associates. https://proceedings.nips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

Lin, S., Hilton, J., y Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods [TruthfulQA: Medición de cómo los modelos imitan falsedades humanas]. En S. Muresan, P. Nakov, y A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3214–3252). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.229

Maynez, J., Narayan, S., Bohnet, B., y McDonald, R. (2020). On faithfulness and factuality in abstractive summarization [Sobre la fidelidad y la exactitud factual en el resumen abstractivo]. En D. Jurafsky, J. Chai, N. Schluter, y J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1906–1919). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.173

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W., Koh, P. W., Iyyer, M., Zettlemoyer, L., y Hajishirzi, H. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long form text generation [FActScore: Evaluación atómica detallada de la precisión factual en la generación de textos extensos]. En H. Bouamor, J. Pino, y K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 12076–12100). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.741

Rajpurkar, P., Jia, R., y Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD [Reconocer lo que no se sabe: Preguntas sin respuesta para SQuAD]. En I. Gurevych y Y. Miyao (Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 784–789). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-2124

