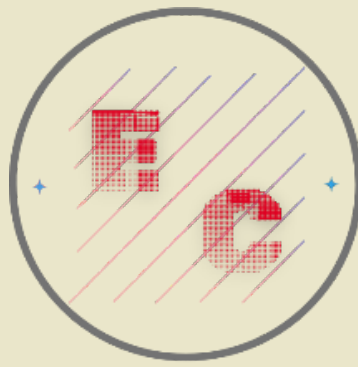




REVISTA MULTIDISCIPLINAR EPISTEMOLOGÍA DE LAS CIENCIAS

Volumen 3, Número 4
Octubre-Diciembre 2026

Edición Trimestral

CROSSREF PREFIX DOI: 10.71112

ISSN: 3061-7812, www.omniscens.com

Revista Multidisciplinar Epistemología de las Ciencias

Volumen 3, Número 4
Octubre-Diciembre 2026

Publicación trimestral
Hecho en México

La Revista Multidisciplinar Epistemología de las Ciencias acepta publicaciones de cualquier área del conocimiento, promoviendo una plataforma inclusiva para la discusión y análisis de los fundamentos epistemológicos en diversas disciplinas. La revista invita a investigadores y profesionales de campos como las ciencias naturales, sociales, humanísticas, tecnológicas y de la salud, entre otros, a contribuir con artículos originales, revisiones, estudios de caso y ensayos teóricos. Con su enfoque multidisciplinario, busca fomentar el diálogo y la reflexión sobre las metodologías, teorías y prácticas que sustentan el avance del conocimiento científico en todas las áreas.

Contacto principal: admin@omniscens.com

Las opiniones expresadas por los autores no necesariamente reflejan la postura del editor de la publicación

Se autoriza la reproducción total o parcial del contenido de la publicación sin previa autorización de la Revista Multidisciplinar Epistemología de las Ciencias siempre y cuando se cite la fuente completa y su dirección electrónica.

Esta obra está bajo una licencia internacional Creative Commons Atribución 4.0.



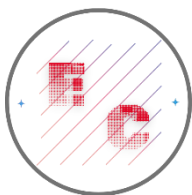
Copyright © 2026: Los autores



9773061781003

Cintillo legal

Revista Multidisciplinar Epistemología de las Ciencias Vol. 3, Núm. 4, Octubre-Diciembre 2026, es una publicación trimestral editada por el Dr. Moises Ake Uc, C. 51 #221 x 16B , Las Brisas, Mérida, Yucatán, México, C.P. 97144 , Tel. 9993556027, Web: <https://www.omniscens.com>, admin@omniscens.com, Editor responsable: Dr. Moises Ake Uc. Reserva de Derechos al Uso Exclusivo No. 04-2024-121717181700-102, ISSN: 3061-7812, ambos otorgados por el Instituto Nacional del Derecho de Autor (INDAUTOR). Responsable de la última actualización de este número, Dr. Moises Ake Uc, fecha de última modificación, 1 octubre 2026.



Revista Multidisciplinar Epistemología de las Ciencias

Volumen 3, Número 4, 2026, octubre-diciembre

DOI: <https://doi.org/10.71112/gpqkd594>

**CUANDO LOS COPILOTOS CONFABULAN: UNA REVISIÓN CRÍTICA DE LAS
ALUCINACIONES EN LA ESCRITURA CIENTÍFICA ASISTIDA POR INTELIGENCIA
ARTIFICIAL**

**WHEN COPILOTS CONFABULATE: A CRITICAL REVIEW OF HALLUCINATIONS IN
AI-ASSISTED SCIENTIFIC WRITING**

Leandro Guerschberg

Yael Estefanía Gutiérrez

Argentina

Cuando los copilotos confabulan: una revisión crítica de las alucinaciones en la escritura científica asistida por inteligencia artificial

When copilots confabulate: a critical review of hallucinations in ai-assisted scientific writing

Leandro Guerschberg^{a,*}

Yael Estefanía Gutiérrez^a

leandro.guerschberg@docentes.unpaz.edu.ar

yael.e.gutierrez@docentes.unpaz.edu.ar

<https://orcid.org/0009-0005-9286-6358>

<https://orcid.org/0009-0003-1616-2995>

*Autor de correspondencia: leandro.guerschberg@docentes.unpaz.edu.ar, ^aUniversidad Nacional de José C. Paz, Argentina

RESUMEN

Antecedentes: Los modelos de lenguaje de gran escala se han incorporado como copilotos en la escritura científica, pero presentan una limitación crítica: la alucinación, entendida como contenido plausible sin sustento factual. Objetivo: Revisar críticamente la frecuencia, tipología, métodos de detección y estrategias de mitigación de las alucinaciones en copilotos de escritura y sus implicaciones para la integridad científica. Métodos: Se realizó una revisión crítica de literatura con búsqueda sistematizada entre enero de 2023 y junio de 2025, organizada según PRISMA 2020; el corpus final quedó constituido por ocho trabajos. Resultados: Los estudios mostraron marcada heterogeneidad según la tarea y la definición operacional: 1,47% de alucinaciones en resúmenes clínicos estructurados, frente a proporciones de referencias inexistentes de 18% a 91,4% en generación abierta según el modelo. Conclusiones: La evidencia no respalda el uso autónomo de copilotos de IA en la producción científica; la verificación humana de las fuentes citadas resulta indispensable.

Palabras clave: modelos de lenguaje de gran escala; alucinación; confabulación; fabricación de referencias; escritura científica; copilotos de IA.

ABSTRACT

Background: Large language models have been increasingly incorporated as copilots in scientific writing, but present a critical limitation: hallucination, understood as plausible content lacking factual support. Objective: To critically review the frequency, typology, detection methods, and mitigation strategies of hallucinations in writing copilots, and their implications for scientific integrity. Methods: A critical literature review with systematized search was conducted between January 2023 and June 2025, organized according to PRISMA 2020; the final corpus comprised eight publications. Results: The studies showed marked heterogeneity according to task and operational definition: 1.47% hallucinations in structured clinical summaries, versus proportions of non-existent references ranging from 18% to 91.4% in open-ended generation depending on the model. Conclusions: The evidence does not support the autonomous use of AI copilots in scientific production; human verification of cited sources remains indispensable. Keywords: large language models; hallucination; confabulation; reference fabrication; scientific writing; AI copilots.

Recibido: 14 de septiembre 2026 | Aceptado: 30 de septiembre 2026 | Publicado: 1 de octubre 2026

INTRODUCCIÓN

Desde el lanzamiento de ChatGPT en noviembre de 2022, los modelos de lenguaje de gran escala (LLM) se incorporaron rápidamente como «copilotos de IA» en los flujos de trabajo de científicos, clínicos y académicos, asistiendo en la redacción, la síntesis de literatura, la sugerencia de marcos metodológicos y la producción de referencias bibliográficas a demanda

(Alkaissi y McFarlane, 2023; Liu et al., 2025). En este trabajo se define un copiloto como un asistente basado en LLM que provee apoyo contextual para tareas de redacción, síntesis y producción de conocimiento en entornos académicos (Lund-Tonnesen et al., 2025; Huang et al., 2025), con independencia de su denominación comercial. Una encuesta global de Nature a investigadores postdoctorales reportó que un tercio de los encuestados usaba IA para mejorar textos, programar u organizar bibliografía (Nordling, 2023).

Esta expansión convive con una vulnerabilidad documentada: la capacidad de los LLM de generar contenido coherente y presentado con seguridad discursiva, pero que puede carecer de sustento factual, fenómeno denominado «alucinación» (Huang et al., 2025), también llamado «alucinación artificial» en el dominio clínico (Alkaissi y McFarlane, 2023) y descrito como un riesgo persistente incluso en aplicaciones críticas (Azamfirei et al., 2023). Farquhar et al. (2024) reservan el término «confabulación» para generaciones incorrectas y arbitrarias cuya respuesta varía ante factores irrelevantes como la semilla aleatoria; Smith et al. (2023) sostienen, en el mismo debate terminológico, que la metáfora neuroanatómica de la confabulación describe con mayor precisión estas generaciones erróneas pero coherentes en su forma. El título de este trabajo emplea «confabular» en su sentido corriente, y el uso técnico se retoma en la sección 3.2.

La distinción no es solo terminológica. Alkaissi y McFarlane (2023) mostraron que ChatGPT, al explicar la patogénesis de la osteoporosis por homocistinuria, produjo una narrativa plausible sin respaldo en la literatura y, al pedir referencias, generó cinco citas con autores e identificadores PubMed de apariencia legítima, ninguno correspondiente a los artículos citados. Walters y Wilder (2023) analizaron 636 citas de GPT-3.5 y GPT-4 en 84 revisiones sobre 42 temas: el 55% (GPT-3.5) y el 18% (GPT-4) correspondían a obras inexistentes, y aun entre citas reales el 43% y el 24%, respectivamente, contenían errores

sustantivos. Chelli et al. (2024), en cirugía ortopédica, reportaron tasas del 39,6% (GPT-3.5), 28,6% (GPT-4) y 91,4% (Bard).

Estas cifras no son mediciones intercambiables de una propiedad estable: dependen de la tarea, la versión del modelo, la unidad de análisis, la definición operacional y las condiciones de despliegue. La variabilidad no reduce, sin embargo, la relevancia del problema: las citas vinculan afirmaciones con evidencia previa (Walters y Wilder, 2023), y una referencia inexistente con apariencia formal correcta es difícil de detectar sin recuperar y contrastar la fuente, con consecuencias que en contextos clínicos exceden la integridad académica y alcanzan la seguridad y la responsabilidad institucional (Asgari et al., 2025; Singhal et al., 2023).

El problema excede la fabricación de referencias: Huang et al. (2025) identifican también contradicciones respecto de una fuente, información no verificable, errores de factualidad y violaciones de fidelidad, mientras que Farquhar et al. (2024) añaden la arbitrariedad como componente adicional. Comprobar que una referencia existe no garantiza que respalde la afirmación asociada, y una respuesta consistente no equivale a evidencia de factualidad.

Los aportes relevantes están dispersos entre campos: taxonomías generales (Huang et al., 2025), fabricación de referencias (Walters y Wilder, 2023; Chelli et al., 2024), riesgos de la escritura académica (Alkaissi y McFarlane, 2023; Liu et al., 2025), detección (Farquhar et al., 2024) y despliegue clínico (Asgari et al., 2025; Singhal et al., 2023). Esta revisión busca integrarlos en torno al uso de los LLM como copilotos de escritura, con cuatro objetivos: (1) caracterizar la tipología y los mecanismos de las alucinaciones; (2) sintetizar la evidencia cuantitativa sobre su frecuencia; (3) examinar los métodos de detección y mitigación; y (4) analizar sus implicaciones para la integridad científica y la gobernanza.

METODOLOGÍA

Diseño y alcance de la revisión

Este trabajo presenta una revisión crítica de literatura con búsqueda bibliográfica sistematizada y síntesis narrativa, orientada a sintetizar evidencia empírica y aportes teóricos sobre las alucinaciones producidas por LLM en la escritura científica y académica; este diseño se justifica por la naturaleza emergente del fenómeno y la heterogeneidad de la literatura disponible (estudios experimentales, evaluaciones comparativas, revisiones de alcance y propuestas taxonómicas).

A diferencia de una revisión sistemática orientada a una estimación cuantitativa agregada, el propósito fue construir un corpus delimitado que permitiera examinar la tipología y los mecanismos de las alucinaciones, su magnitud en distintas tareas y modelos, los métodos de detección y mitigación disponibles, y sus implicaciones para la integridad y la gobernanza, privilegiando la articulación conceptual y la comparación contextualizada antes que la agregación estadística.

Para dar transparencia al proceso de selección se tomaron como referencia las etapas de identificación, cribado, elegibilidad e inclusión del diagrama PRISMA 2020 (Page et al., 2021), usado solo para organizar y comunicar la conformación del corpus, sin metaanálisis, dada la heterogeneidad entre estudios en modelos, tareas, unidades de análisis y definiciones operacionales.

El alcance se concentró en la relación entre estos modos de fallo y la producción de conocimiento científico, con atención a la generación y verificación de referencias, la escritura académica, la síntesis de literatura y la documentación especializada, incorporando también estudios clínicos y biomédicos cuando aportaban elementos pertinentes para comprender los mecanismos o las condiciones que modulan su frecuencia.

Estrategia de búsqueda bibliográfica

La búsqueda se realizó en PubMed/MEDLINE, Scopus, Web of Science Core Collection, ACM Digital Library, Google Scholar, ScienceDirect y JSTOR, así como en recursos de acceso institucional y abierto, combinación justificada por la naturaleza interdisciplinaria del fenómeno (lingüística computacional, IA, recuperación de información, comunicación científica, integridad académica y seguridad clínica).

El período de elegibilidad se delimitó entre enero de 2023 y junio de 2025, coincidiendo con el despliegue masivo de LLM de propósito general utilizables sin mediación técnica especializada; los trabajos anteriores se usaron, cuando fue necesario, solo como antecedentes conceptuales.

Dado que el fenómeno carece de una denominación única (hallucination, confabulation, factuality, faithfulness, fabrication, misinformation), se combinaron con operadores booleanos tres núcleos de términos: el fenómeno de interés; los modelos o sistemas analizados (ChatGPT, GPT, Bard, large language models); y los contextos de aplicación relevantes para la revisión (scientific writing, academic writing, systematic reviews, clinical documentation, biomedical research). Términos específicos vinculados con procedimientos de evaluación, como semantic entropy, benchmark o citation verification, se utilizaron en búsquedas complementarias cuando resultó pertinente. No se restringió el idioma en la búsqueda inicial, aunque los trabajos finalmente incluidos estaban publicados en inglés.

Criterios de inclusión y exclusión

Los criterios de elegibilidad priorizaron la pertinencia del trabajo para el problema analizado y la disponibilidad de información metodológica suficiente, no el prestigio bibliométrico de la revista. Se incluyeron estudios con mediciones empíricas de alucinaciones, fabricación de referencias, errores de factualidad o violaciones de fidelidad, así como trabajos

que, sin una tasa específica, aportaran taxonomías, marcos conceptuales, procedimientos de detección o estrategias de mitigación relevantes.

Se exigió publicación en revistas revisadas por pares con identificación bibliográfica formal. Se excluyeron los preprints sin revisión por pares, los estudios centrados exclusivamente en modelos visión-lenguaje sin relación con la escritura científica, los trabajos anteriores a enero de 2023 y aquellos con pertinencia insuficiente tras el examen del título, el resumen o el texto completo.

La aplicación de estos criterios permitió distinguir evidencia directa (generación de referencias, escritura académica, evaluación de contenido) de evidencia contextual (aportes taxonómicos, técnicos o metodológicos), distinción mantenida durante la síntesis para evitar equiparar hallazgos de tareas metodológicamente diferentes.

Proceso de selección de los estudios

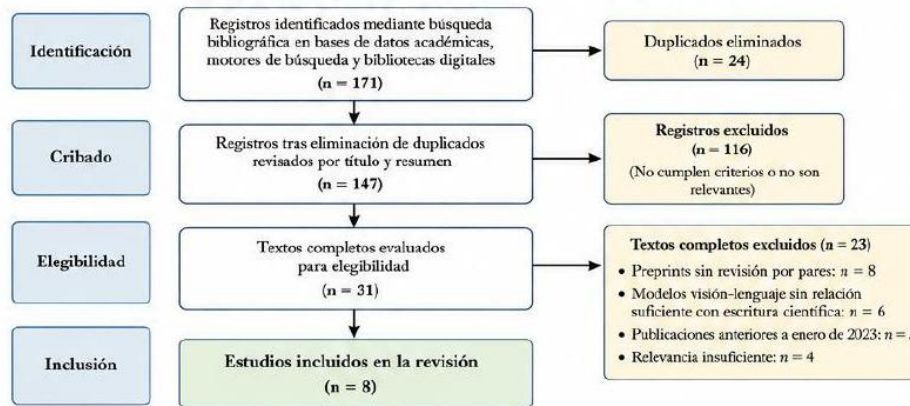
El proceso de selección se organizó en cuatro etapas siguiendo PRISMA 2020 (Page et al., 2021); el cribado por título y resumen y la elegibilidad a texto completo se realizaron de forma dual e independiente por los dos autores (Guerschberg y Gutiérrez), resolviendo discrepancias por consenso.

Se identificaron 171 registros; tras eliminar 24 duplicados quedaron 147 para cribado, etapa en la que se excluyeron 116 registros sin relación suficiente con el problema. Los 31 trabajos restantes se evaluaron a texto completo; se excluyeron 23 (ocho preprints, seis sobre modelos visión-lenguaje, cinco anteriores a 2023 y cuatro con pertinencia insuficiente), quedando ocho trabajos en el corpus final, con diseños y finalidades diferentes conservados deliberadamente por su valor complementario.

El proceso completo se resume en la Figura 1.

Figura 1

Diagrama de flujo PRISMA del proceso de selección de trabajos



Fuente: elaboración propia a partir de la estructura de PRISMA 2020 (Page et al., 2021).

Estrategia de análisis y síntesis

Los ocho trabajos incluidos se examinaron mediante una estrategia narrativa y comparativa orientada a identificar regularidades, diferencias y relaciones entre los hallazgos, preservando las condiciones bajo las cuales cada resultado fue producido, considerando autores, año, revista, tipo de estudio, muestra, modelo, tarea, procedimiento de medición y resultados principales.

La interpretación prestó atención central a la definición operacional de alucinación de cada trabajo, dado que el término no designa una única clase de error; las tasas se interpretaron en relación con la definición, la unidad de análisis y el procedimiento de verificación de cada investigación, y no como estimaciones equivalentes de una única variable, con una finalidad analítica y contextual antes que metaanalítica.

La síntesis se estructuró en cuatro ejes: (1) tipología y mecanismos de las alucinaciones (intrínsecas/extrínsecas, factualidad/fidelidad, confabulación); (2) frecuencia y tasas cuantitativas, con atención a la fabricación de referencias; (3) métodos de detección y

mitigación; y (4) implicaciones para la integridad científica y la gobernanza. Se distinguió, como criterio transversal, entre evidencia directa y contextual, para evitar la extrapolación automática entre tareas metodológicamente diferentes.

Conviene distinguir, para la lectura de las secciones siguientes, entre el corpus analítico principal (los ocho trabajos PRISMA, fuente exclusiva de los resultados de la Sección 3) y la literatura contextual complementaria (fuentes adicionales verificables citadas en la Introducción y la Discusión para situar el problema, sin integrar el corpus ni aportar resultados propios de esta revisión).

RESULTADOS

Panorama de los estudios incluidos

El corpus final estuvo constituido por ocho trabajos publicados entre 2023 y 2025, con marcada heterogeneidad en diseños, unidades de análisis, modelos y contextos de aplicación, lo que permitió integrar evidencia directamente vinculada con la escritura científica junto con aportes conceptuales y metodológicos de otros escenarios de uso de los LLM.

Cuatro trabajos aportaron evidencia directa: Walters y Wilder (2023) y Chelli et al. (2024) evaluaron la fabricación y precisión de referencias; Alkaissi y McFarlane (2023) documentaron tempranamente afirmaciones y referencias inexistentes; y Liu et al. (2025) examinaron oportunidades, límites y problemas éticos del uso de ChatGPT en escritura académica. Los cuatro restantes aportaron evidencia contextual: Huang et al. (2025) desarrollaron una taxonomía amplia; Farquhar et al. (2024) propusieron un método de detección por entropía semántica; Singhal et al. (2023) evaluaron un modelo médico especializado; y Asgari et al. (2025) analizaron tasas de alucinación y omisión en un marco clínico estructurado. Lund-Tonnesen et al. (2025), citado en la introducción, no integra este corpus; se usa solo como antecedente conceptual.

La Tabla 1 resume las características de los ocho trabajos incluidos; la distinción entre evidencia directa y contextual es analítica y no implica jerarquización de calidad.

Tabla 1.

Resumen de los trabajos incluidos en la revisión crítica (2023-2025)

Autor(es)	Año	Revista	Tipo de estudio	N / corpus	Aporte principal a la revisión
Alkaissi y McFarlane	2023	Cureus	Reporte de caso / editorial	2 temas	Documentación temprana de afirmaciones y referencias bibliográficas inexistentes
Huang et al.	2025	ACM Transactions on Information Systems	Revisión sistemática / encuesta del campo	Revisión extensa	Taxonomía de alucinaciones intrínsecas y extrínsecas; facticidad y fidelidad
Liu et al.	2025	Arquivos Brasileiros de Oftalmologia	Revisión de alcance	327 documentos	Riesgos y desafíos de ChatGPT en escritura académica
Chelli et al.	2024	Journal of Medical Internet Research	Estudio empírico comparativo	471 referencias; 33 consultas	Tasas de fabricación y precisión de referencias en GPT-3.5, GPT-4 y Bard
Singhal et al.	2023	Nature	Evaluación de benchmark	MultiMedQA	Desempeño de Med-PaLM y persistencia de limitaciones de facticidad y seguridad
Farquhar et al.	2024	Nature	Estudio metodológico experimental	Múltiples LLM y tareas de pregunta-respuesta	Detección de confabulaciones mediante entropía semántica
Walters y Wilder	2023	Scientific Reports	Estudio empírico	636 citas; 84 revisiones	Fabricación y errores en referencias generadas por GPT-3.5 y GPT-4
Asgari et al.	2025	npj Digital Medicine	Marco metodológico + estudio empírico	12.999 oraciones	Tasas de alucinación y omisión bajo condiciones clínicas estructuradas

Fuente: elaboración propia. Nota: N refiere al tamaño muestral o a las características del corpus según el diseño de cada trabajo.

Taxonomía y mecanismos de las alucinaciones en copilotos de IA

Huang et al. (2025) proporcionan el marco taxonómico más amplio del corpus. Las alucinaciones intrínsecas contradicen información verificable de la fuente (por ejemplo, invertir la conclusión de un artículo citado); las extrínsecas incorporan información no verificable a partir de la fuente o de registros externos, de lo cual la fabricación de una referencia inexistente es el ejemplo más relevante. En una segunda dimensión, los errores de factualidad implican información empíricamente incorrecta, mientras que las violaciones de fidelidad distorsionan el contenido de una fuente existente.

Farquhar et al. (2024) incorporan el concepto de confabulación para respuestas incorrectas y arbitrarias: ante consultas equivalentes, el modelo puede producir respuestas semánticamente divergentes (en su ejemplo, sobre el blanco molecular de Sotorasib, KRAS-G12C correcto frente a KRAS-G12D incorrecto), problema de naturaleza distinta a un error consistente y que exige mecanismos de detección propios.

Asgari et al. (2025) distinguen además alucinaciones (información no respaldada por la fuente) de omisiones (información relevante no incorporada): sobre 12.999 oraciones evaluadas, reportaron 3,45% de omisión y 1,47% de alucinación, mostrando que evaluar la calidad de una salida no puede limitarse a la información falsa incorporada. En conjunto, «alucinación» agrupa modos de fallo diferentes: verificar la existencia de una referencia no garantiza fidelidad, ni la consistencia garantiza factualidad.

Frecuencia y tasas cuantitativas de alucinación

Las tasas de alucinación del corpus mostraron variabilidad considerable, coherente con la advertencia metodológica de la sección 2.5: las cifras corresponden a modelos, tareas y definiciones operacionales diferentes.

Citas bibliográficas fabricadas

Walters y Wilder (2023), sobre 636 citas de GPT-3.5 y GPT-4 en 84 revisiones de 42 temas, hallaron que el 55% (GPT-3.5) y el 18% (GPT-4) de las referencias correspondían a obras inexistentes, y que entre las citas reales el 43% y el 24%, respectivamente, contenían al menos un error sustantivo en autoría, título, año, revista o paginación: el problema comprende, por tanto, tanto la fabricación completa como la alteración parcial de una fuente real. Esto es consistente con la naturaleza generativa de los LLM, que en ausencia de recuperación externa combinan nombres, títulos y años de manera verosímil sin corresponder a un registro bibliográfico real.

Chelli et al. (2024), sobre 33 consultas y 471 referencias para once revisiones sistemáticas en cirugía ortopédica, reportaron tasas de alucinación del 39,6% (GPT-3.5), 28,6% (GPT-4) y 91,4% (Bard), con precisión en recuperar referencias reales del 9,4%, 13,4% y 0% respectivamente, desempeño insuficiente para considerar estos modelos sustitutos autónomos de una estrategia bibliográfica formal.

El caso de Alkaissi y McFarlane (2023) ilustra la dimensión cualitativa del problema: el PMID 12352394, ofrecido como respaldo de una explicación sobre homocisteína y osteocalcina, correspondía en realidad a un trabajo sobre grapas de titanio en cirugía urológica, falsedad detectable solo mediante recuperación efectiva del documento. En conjunto, estos estudios muestran que la fabricación bibliográfica puede producir una fuente inexistente, una referencia parcialmente incorrecta o una atribución formalmente válida pero desvinculada de la afirmación que pretende respaldar.

Alucinaciones en la síntesis de conocimiento médico

Singhal et al. (2023) desarrollaron Med-PaLM, que alcanzó 67,6% de precisión en MedQA, pero la evaluación humana mostró que ese desempeño no eliminaba errores de factualidad, razonamiento y seguridad, particularmente ante razonamiento causal o integración de información de distintos dominios clínicos.

Este resultado es relevante como evidencia contextual: un desempeño elevado en benchmarks no garantiza confiabilidad epistémica ni permite estimar una tasa de alucinación en escritura científica.

Alucinaciones bajo condiciones de despliegue clínico estructurado

Asgari et al. (2025) reportaron 1,47% de alucinación en resumen clínico restringido, la tasa más baja del corpus, con taxonomía explícita de errores, 18 configuraciones experimentales, evaluación de seguridad y anotación humana (CREOLA) sobre documentos fuente disponibles como contexto, junto con 3,45% de omisión. Esto sugiere que condiciones estructuradas, fuentes explícitas y validación pueden reducir ciertos errores, aunque el contraste con Walters y Wilder (2023) o Chelli et al. (2024) debe leerse como ilustración de la amplitud de resultados posibles, no como efecto atribuible a una arquitectura específica.

Factores asociados con la variabilidad de las tasas de alucinación

La comparación contextual de los estudios permite identificar patrones emergentes, no efectos causales aislados, dada la heterogeneidad del corpus. La versión del modelo es el factor más consistente: tanto Walters y Wilder (2023) como Chelli et al. (2024) hallaron menores tasas en GPT-4 que en GPT-3.5, y Bard presentó una tasa de alucinación sensiblemente mayor que ambas versiones de GPT.

La estructura de la tarea y la disponibilidad de fuentes constituyen un segundo factor: la generación abierta de referencias exige producir información sin documentos de contexto, mientras que el resumen restringido provee información explícita; los resultados son

compatibles con la hipótesis de que un mayor anclaje documental reduce ciertas alucinaciones, sin que el corpus permita cuantificar este efecto de forma aislada.

La densidad terminológica del dominio puede aumentar la plausibilidad formal de un error (por ejemplo, en literatura biomédica), y la estructura formal de las referencias genera un perfil de riesgo particular: verificar una cita exige comprobación externa, de modo que el costo de verificación supera ampliamente el de producción, generando una asimetría entre facilidad de generación y dificultad de detección.

Métodos de detección y aproximaciones de control

Farquhar et al. (2024) proponen la entropía semántica, que estima la incertidumbre a nivel del significado generando múltiples respuestas para una consulta, agrupándolas por equivalencia semántica y midiendo su dispersión: una entropía elevada indica respuestas divergentes, comportamiento propio de la confabulación. El método no depende de la forma superficial del texto, pero no identifica errores factuales sistemáticos y consistentes, y aumenta el costo computacional.

Asgari et al. (2025) representan una aproximación complementaria basada en la estructuración integral del proceso (taxonomías de error, anotación humana, evaluación de seguridad, refinamiento iterativo), mostrando que la reducción de alucinaciones depende también del diseño del flujo de trabajo, no solo de la arquitectura del modelo.

Ningún procedimiento único detecta todas las formas de alucinación: confabulaciones, errores factuales sistemáticos, referencias inexistentes y violaciones de fidelidad requieren mecanismos de control diferenciados, combinando detección automática con verificación externa y revisión humana.

DISCUSIÓN

El copiloto no confiable: una síntesis crítica

La evidencia integrada converge en que los LLM no ofrecen, en las condiciones estudiadas, garantías suficientes para ser fuentes autónomas de contenido factual o bibliográfico; su fluidez y velocidad explican su atractivo como herramientas de productividad, pero dificultan identificar errores cuando el contenido conserva apariencia de plausibilidad académica. Esta organización de las alucinaciones sistematizada por Huang et al. (2025) es, además, consistente con marcos más generales de la generación de lenguaje natural, como Ji et al. (2023), literatura contextual ajena al corpus.

La alucinación no es una propiedad binaria ni una tasa estable atribuible a un modelo con independencia del contexto: varía según la versión, la tarea, la disponibilidad de fuentes, la definición operacional y el procedimiento de verificación, sin permitir una escala lineal entre condiciones de despliegue, aunque sí advertir que la confiabilidad de una respuesta depende de la interacción entre modelo, tarea y condiciones de uso. En la misma línea, aunque fuera del corpus analítico, Anh-Hoang et al. (2025) atribuyen parte de las alucinaciones reportadas en la literatura a estrategias de prompting subóptimas y no solo al comportamiento intrínseco del modelo.

El problema no reside solo en el error sino en su presentación: las referencias fabricadas conservan autores plausibles, revistas reconocibles y títulos verosímiles, sin una señal discursiva que distinga la generación correcta de la fabricación, por lo que el usuario debe recurrir a verificación externa. El concepto de copiloto es aquí pertinente: estos sistemas asisten en redacción y exploración, pero la evidencia no respalda transferirles la responsabilidad epistémica, que permanece en quienes utilizan y firman el contenido.

Implicaciones para la integridad científica

La integridad de la comunicación científica depende de la atribución precisa de fuentes, la verificabilidad, la transparencia metodológica y la responsabilidad sobre el contenido publicado; las alucinaciones de los LLM no son solo un problema técnico, sino que pueden alterar la relación entre una afirmación y la evidencia que pretende sostenerla.

Las referencias fabricadas son el caso más evidente: introducen evidencia irrecuperable y atribuyen autoridad a autores que no produjeron el conocimiento citado. Las violaciones de fidelidad son menos visibles, pues la existencia efectiva del artículo genera una apariencia de validación, aunque su contenido haya sido tergiversado; comprobar que una referencia existe y comprobar lo que efectivamente afirma son operaciones distintas. Fuera del corpus, Bhattacharyya et al. (2023) refuerzan este patrón: de 115 referencias médicas de ChatGPT, 47% eran fabricadas y 46% reales pero inexactas, y solo 7% resultó auténtico y preciso. Esto importa especialmente cuando los LLM aceleran revisiones de literatura: si la verificación posterior no ocurre, el ahorro de tiempo reduce la trazabilidad epistémica del manuscrito. Liu et al. (2025), en 327 documentos, identificaron además información desactualizada, riesgo de plagio, sesgos, ambigüedad de autoría y erosión de la escritura como proceso de aprendizaje. Un LLM no puede asumir responsabilidad por una afirmación falsa ni garantizar que sus referencias fueron verificadas; esa responsabilidad permanece en los autores humanos, de modo que el copiloto no reduce sus obligaciones tradicionales, sino que agrega obligaciones de comprobación y transparencia. En esta línea, el Committee on Publication Ethics (2023) sostiene que un LLM no puede figurar como autor porque no puede asumir responsabilidad por el contenido ni declarar conflictos de interés, y que su uso debe declararse explícitamente en el manuscrito. La declaración del uso de IA debe, por tanto, hacer visible en qué etapa intervino la herramienta y qué se verificó.

Estrategias de mitigación

Las estrategias disponibles para reducir el riesgo operan en niveles distintos del problema, sin que la heterogeneidad del corpus permita establecer una jerarquía definitiva entre ellas: modifican la información disponible, estructuran la generación, detectan incertidumbre o desplazan el control hacia la verificación humana.

Anclar la generación en fuentes explícitas, por provisión directa o por RAG, es una primera vía; el corpus no incluye estudios que aislen el efecto de RAG, pero la tasa de 1,47% de Asgari et al. (2025) con documentos fuente disponibles, contrastada con las tasas superiores de Chelli et al. (2024) en generación abierta, ilustra la importancia potencial del anclaje documental sin permitir establecer causalidad. Fuera del período de elegibilidad, antecedentes en generación conversacional señalan que la recuperación documental puede reducir el contenido no sustentado (Shuster et al., 2021). La disponibilidad de fuentes no elimina todo error: un sistema puede representar incorrectamente una fuente pertinente.

Los marcos de despliegue estructurado son una segunda vía: Asgari et al. (2025) combinaron taxonomías de error, configuraciones alternativas, evaluación de seguridad y anotación humana, mostrando que la reducción del riesgo depende también del diseño del flujo de trabajo. La verificación independiente de referencias, tercera vía y la de aplicación más inmediata para el autor individual, exige contrastar toda referencia propuesta con una fuente externa, comprobando existencia, datos bibliográficos y correspondencia con la afirmación citada, precaución justificada por las tasas de Walters y Wilder (2023) y Chelli et al. (2024).

La cuarta vía es la detección automática: Farquhar et al. (2024) muestran que la entropía semántica identifica confabulaciones por variabilidad semántica, aunque no detecta errores factuales sistemáticos y consistentes. Fuera del corpus, Manakul et al. (2023) proponen un método zero-resource similar (SelfCheckGPT) basado en la consistencia interna de

múltiples respuestas. La Tabla 2 sintetiza las cuatro vías identificadas en el corpus y su principal limitación; ninguna resuelve el problema de manera independiente.

Tabla 2.

Síntesis de estrategias de mitigación identificadas en el corpus

Estrategia	Nivel del problema sobre el que actúa	Evidencia en el corpus	Principal limitación
Provisión o recuperación de información fuente	Modifica la información disponible para el modelo antes o durante la generación	Asgari et al. (2025): generación estructurada con documentos fuente disponibles; evidencia contextual sobre la relevancia del anclaje documental, sin evaluación específica de RAG	No garantiza la fidelidad de la interpretación y el corpus no permite estimar de manera aislada el efecto de la recuperación documental
Marcos de despliegue estructurado	Estructura el proceso de generación y evaluación	Asgari et al. (2025): taxonomías de error, configuraciones múltiples, anotación humana	Requiere rediseño del flujo de trabajo, no solo del modelo
Verificación independiente de referencias	Desplaza el control hacia la revisión humana posterior	Justificada por las tasas de fabricación de Walters y Wilder (2023) y Chelli et al. (2024)	Depende enteramente de la diligencia del autor; no es automatizable sin herramientas externas
Detección automática mediante entropía semántica	Identifica incertidumbre en el momento de la generación	Farquhar et al. (2024): detección de confabulaciones por variabilidad semántica	No detecta errores factuales sistemáticos y consistentes; aumenta el costo computacional

Fuente: elaboración propia a partir del análisis narrativo del corpus.

Dimensiones de gobernanza

La incorporación de copilotos de IA plantea problemas que exceden las decisiones individuales de los investigadores y justifican dimensiones institucionales de gobernanza que involucren revistas, universidades, organismos de investigación y desarrolladores.

La transparencia sobre el uso de IA debería diferenciar entre corrección lingüística, reformulación, síntesis de documentos propios, generación inicial de contenido y propuesta de referencias, tareas con perfiles de riesgo distintos. Fuera del corpus, Yoo (2025) compara políticas editoriales internacionales y encuentra convergencia parcial en la prohibición de autoría no humana, con variaciones en el detalle exigido; Ganjavi et al. (2024), en cien editoriales y cien revistas líderes, hallaron que solo 17% de los editores y 70% de las revistas ofrecían pautas específicas sobre IA generativa. Se requieren además instrumentos estandarizados: los diseños de Walters y Wilder (2023) y Chelli et al. (2024), estos sí del corpus, muestran que es posible operacionalizar la fabricación y precisión de referencias, aunque la diversidad de criterios dificulta comparar modelos en el tiempo.

Los usuarios tampoco disponen de información interpretable sobre el desempeño esperado por tarea, ya que un benchmark general no permite inferir el comportamiento en generación de referencias o síntesis de literatura. La gobernanza debe distribuir responsabilidades: desarrolladores, instituciones y revistas deben reducir el riesgo previsible y definir condiciones de uso y verificación, mientras que los autores mantienen la responsabilidad final sobre lo que incorporan al registro científico.

Limitaciones de esta revisión

Esta revisión presenta limitaciones. Se trata de una revisión crítica con un corpus final de ocho trabajos, que no pretende ser un relevamiento exhaustivo de la literatura sobre alucinaciones en LLM; pueden existir investigaciones pertinentes no incorporadas. El corpus combina evidencia directa sobre escritura científica con evidencia contextual clínica y metodológica, lo que limita la extrapolación directa de ciertos resultados a los procesos de escritura académica.

La heterogeneidad entre modelos, versiones, tareas, unidades de análisis y procedimientos de verificación impide interpretar las tasas como mediciones de una única variable homogénea, por lo que se adoptó una estrategia narrativa y no un metaanálisis. Las tasas cuantitativas, además, corresponden a versiones específicas de modelos en rápida evolución y no deben interpretarse como propiedades permanentes; la delimitación temporal (enero de 2023 a junio de 2025) excluyó del corpus principal trabajos teóricos anteriores, usados solo como contexto conceptual.

Los ocho trabajos incluidos estaban publicados en inglés, por lo que la revisión no permite establecer si el fenómeno se comporta de modo diferente en otros idiomas. Tampoco se aplicó una herramienta estandarizada de evaluación del riesgo de sesgo; los estudios se examinaron críticamente con una finalidad interpretativa, coherente con el carácter crítico y la estrategia de síntesis narrativa adoptada, lo que debe considerarse al ponderar la fortaleza de las inferencias derivadas de un corpus pequeño y heterogéneo.

CONCLUSIONES

La evidencia integrada muestra de manera consistente que los copilotos de escritura basados en IA no ofrecen, en las condiciones examinadas, garantías suficientes para ser herramientas autónomas de producción de conocimiento científico. Los estudios centrados en generación bibliográfica documentaron proporciones relevantes de referencias inexistentes o incorrectas, variables según el modelo, la tarea y la definición operacional; estas cifras no representan una tasa universal de alucinación, sino evidencia de que, en ciertos escenarios de escritura y búsqueda bibliográfica, el riesgo alcanza magnitudes incompatibles con la aceptación no verificada del contenido generado.

La comprensión conceptual del fenómeno también avanzó: la distinción entre alucinaciones intrínsecas y extrínsecas, junto con los ejes de factualidad y fidelidad (Huang et al., 2025) y la noción técnica de confabulación (Farquhar et al., 2024), permite diferenciar modos de fallo con consecuencias distintas para la escritura científica. Asgari et al. (2025) muestran, desde otro contexto, que condiciones estructuradas de generación y mecanismos sistemáticos de evaluación pueden acompañarse de tasas reducidas, sin que ello demuestre que una única estrategia elimine el problema.

De la síntesis emergen tres implicaciones. Primero, toda referencia propuesta por un modelo debe verificarse de manera independiente, no solo en su existencia sino en la exactitud de sus datos y en la correspondencia entre la fuente y la afirmación que respalda; delegar esta responsabilidad en el sistema generativo introduce un riesgo injustificable donde la trazabilidad de las fuentes es condición básica de integridad. Segundo, el uso de IA en la preparación de manuscritos requiere niveles de transparencia proporcionales al tipo de intervención, de modo que la comunidad científica pueda evaluar las condiciones bajo las cuales fue construido el texto. Tercero, resulta necesario avanzar hacia instrumentos de evaluación más comparables y

específicos para los usos científicos de los LLM, dado que la heterogeneidad actual dificulta establecer comparaciones estables entre modelos y contextos.

La integración de copilotos de IA en la escritura científica ya forma parte del escenario contemporáneo de producción de conocimiento. El problema central no es determinar si estas herramientas serán utilizadas, sino bajo qué condiciones de control, responsabilidad y verificación podrán incorporarse sin debilitar la integridad del registro científico. El copiloto puede ser diligente, veloz y elocuente, pero también puede fabricar, distorsionar o confabular sin que esas fallas resulten evidentes en la superficie del texto; reconocer esa tensión, y diseñar prácticas capaces de gestionarla, es un paso necesario hacia una ciencia asistida por inteligencia artificial productiva, verificable y confiable.

Declaración de conflicto de interés

Los autores declaran no tener ningún conflicto de interés relacionado con esta investigación.

Declaración de contribución a la autoría

Leandro Guerschberg: conceptualización, metodología, investigación (criado por título y resumen y evaluación de elegibilidad a texto completo, realizados de forma independiente), análisis formal, redacción del borrador original, revisión y edición de la redacción, visualización, supervisión y administración del proyecto

Yael Estefanía Gutiérrez: investigación (criado por título y resumen y evaluación de elegibilidad a texto completo, realizados de forma independiente), curación de datos, revisión y edición de la redacción.

Declaración de uso de inteligencia artificial

Durante la preparación del presente manuscrito se utilizaron herramientas de inteligencia artificial generativa como apoyo editorial: ChatGPT, modelo GPT-5.6 Sol (OpenAI), y Claude Sonnet 5.0, configuración ALTO (Anthropic). Su utilización se limitó a tareas de corrección de estilo y sintaxis, mejora de la claridad y fluidez de la redacción, reducción del apartado Resumen y revisión y optimización de las traducciones al idioma inglés. Estas herramientas no fueron utilizadas para la generación autónoma de datos, resultados, análisis ni conclusiones de la investigación. Todo contenido producido, revisado o modificado con asistencia de inteligencia artificial fue posteriormente verificado, evaluado y aprobado por los autores, quienes asumen plena responsabilidad por la exactitud, integridad y contenido final del manuscrito.

REFERENCIAS

- Alkaissi, H., y McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*, 15(2), e35179. <https://doi.org/10.7759/cureus.35179>
- Anh-Hoang, D., Tran, V., y Nguyen, L.-M. (2025). Survey and analysis of hallucinations in large language models: Attribution to prompting strategies or model behavior. *Frontiers in Artificial Intelligence*, 8, 1622292. <https://doi.org/10.3389/frai.2025.1622292>
- Asgari, E., Montana-Brown, N., Dubois, M., Khalil, S., Balloch, J., Au Yeung, J., y Pimenta, D. (2025). A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *npj Digital Medicine*, 8, 274. <https://doi.org/10.1038/s41746-025-01670-7>
- Azamfirei, R., Kudchadkar, S. R., y Fackler, J. (2023). Large language models and the perils of their hallucinations. *Critical Care*, 27, 120. <https://doi.org/10.1186/s13054-023-04393-x>

- Bhattacharyya, M., Miller, V. M., Bhattacharyya, D., y Miller, L. E. (2023). High rates of fabricated and inaccurate references in ChatGPT-generated medical content. *Cureus*, 15(5), e39238. <https://doi.org/10.7759/cureus.39238>
- Chelli, M., Descamps, J., Lavoue, V., Trojani, C., Azar, M., Deckert, M., Raynier, J.-L., Clowez, G., Boileau, P., y Ruetsch-Chelli, C. (2024). Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: Comparative analysis. *Journal of Medical Internet Research*, 26, e53164. <https://doi.org/10.2196/53164>
- Committee on Publication Ethics. (2023). Authorship and AI tools [Declaración de posición]. <https://doi.org/10.24318/cCVRZBms>
- Farquhar, S., Kossen, J., Kuhn, L., y Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, 630, 625-630. <https://doi.org/10.1038/s41586-024-07421-0>
- Ganjavi, C., Eppler, M. B., Pekcan, A., Biedermann, B., Abreu, A., Collins, G. S., Gill, I. S., y Cacciamani, G. E. (2024). Publishers' and journals' instructions to authors on use of generative artificial intelligence in academic and scientific publishing: Bibliometric analysis. *BMJ*, 384, e077192. <https://doi.org/10.1136/bmj-2023-077192>
- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., y Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), Artículo 42. <https://doi.org/10.1145/3703155>
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., y Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), Artículo 248. <https://doi.org/10.1145/3571730>

- Liu, Y., Kong, W., y Merve, K. (2025). ChatGPT applications in academic writing: A review of potential, limitations, and ethical challenges. *Arquivos Brasileiros de Oftalmologia*, 88(3), e2024-0269. <https://doi.org/10.5935/0004-2749.2024-0269>
- Lund-Tonnesen, M., Vahr Lauridsen, S., y Rosenberg, J. (2025). Evaluating Microsoft Copilot in Qualitative Health Research: Accurate for Manifest Content Coding but Limited in Latent Interpretation. *Cureus*, 17(10), e95719. <https://doi.org/10.7759/cureus.95719>
- Manakul, P., Liusie, A., y Gales, M. J. F. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 9004-9017. <https://doi.org/10.18653/v1/2023.emnlp-main.557>
- Nordling, L. (2023). How ChatGPT is transforming the postdoc experience. *Nature*, 622(7983), 655-657. <https://doi.org/10.1038/d41586-023-03235-8>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., y Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- Shuster, K., Poff, S., Chen, M., Kiela, D., y Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3784-3803. <https://doi.org/10.18653/v1/2021.findings-emnlp.320>
- Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., ... Natarajan, V. (2023).

Large language models encode clinical knowledge. *Nature*, 620, 172-180.

<https://doi.org/10.1038/s41586-023-06291-2>

Smith, A. L., Greaves, F., y Panch, T. (2023). Hallucination or confabulation? Neuroanatomy as metaphor in large language models. *PLOS Digital Health*, 2(11), e0000388.

<https://doi.org/10.1371/journal.pdig.0000388>

Walters, W. H., y Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*, 13, 14045. <https://doi.org/10.1038/s41598-023-41032-5>

Yoo, J. H. (2025). Defining the boundaries of AI use in scientific writing: A comparative review of editorial policies. *Journal of Korean Medical Science*, 40(23), e187.

<https://doi.org/10.3346/jkms.2025.40.e187>

Anexo. Estrategia de búsqueda reproducible (material suplementario)

Con el propósito de facilitar la reproducibilidad del proceso descrito en la Sección 2.2, se detalla a continuación la estrategia de búsqueda utilizada. La última ejecución de las búsquedas en todas las bases y plataformas indicadas se realizó el 3 de agosto de 2025.

Ecuación utilizada en PubMed/MEDLINE: (hallucination[Title/Abstract] OR confabulation[Title/Abstract] OR "reference fabrication"[Title/Abstract] OR factuality[Title/Abstract] OR faithfulness[Title/Abstract] OR misinformation[Title/Abstract]) AND ("large language model"[Title/Abstract] OR ChatGPT[Title/Abstract] OR GPT[Title/Abstract] OR Bard[Title/Abstract] OR "AI writing assistant"[Title/Abstract]) AND ("scientific writing"[Title/Abstract] OR "academic writing"[Title/Abstract] OR "systematic review"[Title/Abstract] OR "clinical documentation"[Title/Abstract] OR "biomedical research"[Title/Abstract])

Esta ecuación se adaptó a la sintaxis específica de cada plataforma (Scopus, Web of Science Core Collection, ACM Digital Library, Google Scholar, ScienceDirect y JSTOR), conservando los tres núcleos conceptuales descritos en la Sección 2.2 —fenómeno de interés, modelos o sistemas analizados y contextos de aplicación— y ajustando operadores booleanos y truncamientos según las capacidades de cada motor de búsqueda. Los términos vinculados con procedimientos específicos de evaluación se utilizaron en búsquedas complementarias cuando resultó pertinente. El desglose del número de registros aportado individualmente por cada base de datos no fue conservado de forma sistemática durante la búsqueda original y constituye una limitación para la reproducibilidad detallada de la estrategia de búsqueda. El número total de registros identificados (171) y el proceso posterior de cribado se documentan en la Sección 2.4 y en la Figura 1.