bid

Revista científica editada por la Facultad de Información y Medios Audiovisuales de la Universidad de Barcelona
ISSN 1575-5886

BiD 56 (junio 2026)
DOI: 10.1344/bid2026.56.02

Evaluación de ChatGPT para la corrección de referencias bibliográficas en estilo Vancouver

Evaluation of ChatGPT for Correcting Bibliographic References in Vancouver Style

Avaluació de ChatGPT per a la correcció de referències bibliogràfiques en estil Vancouver

José-Luis Llopis-Agelán
Profesor jubilado. Universidad Complutense de Madrid.
jlllopis@ucm.es

José-Manuel Estrada-Lorenzo
Biblioteca. Hospital Universitario del Sureste.
josemanuel.estrada@salud.madrid.org

Óliver Martín-Martín
Universidad Complutense de Madrid
omartinm@ucm.es

Ramón Del-Gallego-Lastra
Universidad Complutense de Madrid.
rgallego@ucm.es

Recibido: 03-12-2025
Aceptado: 26-04-2026

Citación recomendada

Llopis-Agelán, José-Luis; Estrada-Lorenzo, José-Manuel; Martín-Martín, Óliver; Del-Gallego-Lastra, Ramón (2026). Evaluación de ChatGPT para la corrección de referencias bibliográficas en estilo Vancouver. *BiD*, 56. https://doi.org/10.1344/bid2026.56.02

Resumen

Objetivos: Evaluar la utilidad y precisión de ChatGPT-4 Turbo para corregir referencias de artículos de revista según el estilo de citación Vancouver, identificando su comportamiento y limitaciones.

Metodología: Estudio descriptivo con 140 referencias de artículos de revista extraídas de 20 Trabajos de Fin de Grado de Enfermería de la Universidad Complutense de Madrid (2019-2020). Se elaboró una lista de referencias patrón mediante revisión por pares. Se solicitó a ChatGPT-4 Turbo la corrección y formateo de las referencias originales usando un *prompt* básico. Su respuesta se comparó con el patrón mediante un sistema categorizado por seis códigos y el recuento de acciones por elemento bibliográfico (n = 521).

Resultados: El 49,3 % de las referencias recibieron ≥1 intervención, mientras que el 50,7 % permanecieron sin cambios; entre estas últimas, 12 ya eran correctas y 59 seguían conteniendo errores. De las 521 acciones, 130 (25,0 %) fueron adecuadas y 391 (75,0 %) inadecuadas (sin intervención 79,3 %, corrección errónea 13,0 %, error de formato 7,2 % y alucinación 0,5 %).

Conclusiones: Con un *prompt* elemental, ChatGPT-4 Turbo genera mejoras limitadas y predomina la ausencia de intervención. Su uso podría ser útil dentro de un protocolo (*prompts* específicos, verificación de metadatos mediante DOI/PMID y supervisión humana), especialmente en fases iniciales o como apoyo previo a la revisión experta.

Palabras clave

Inteligencia artificial; evaluación; referencias bibliográficas; estilo Vancouver.

Abstract

Objectives: To assess the usefulness and accuracy of ChatGPT-4 Turbo in correcting journal article references according to the Vancouver style, and to characterize its behavior and limitations.

Methodology: Descriptive study including 140 journal article references extracted from 20 undergraduate nursing theses at the Complutense University of Madrid (2019-2020). A gold-standard reference list was created via peer review. ChatGPT-4 Turbo was asked to correct and format the original references using a basic prompt. Its output was compared with the gold-standard using a six-category coding system and by counting actions per bibliographic element (n = 521).

Results: Overall, 49.3 % of references received ≥1 intervention, whereas 50.7 % remained unchanged; among the latter, 12 were already correct and 59 still contained errors. Of 521 actions, 130 (25.0 %) were adequate and 391 (75.0 %) inadequate (non-intervention 79.3 %, erroneous correction 13.0 %, formatting error 7.2 %, hallucination 0.5 %).

Conclusions: With a basic prompt, ChatGPT-4 Turbo yields limited improvements and non-intervention predominates. Its application could be useful within a protocol (task-specific prompts, DOI/PMID-based metadata verification, human oversight), particularly at early stages or as support prior to expert review.

Keywords

Artificial intelligence; assessment; bibliographic references; Vancouver style.

Resum

Objectius: Avaluar la utilitat i la precisió de ChatGPT-4 Turbo en la correcció de referències d'articles de revista segons l'estil Vancouver, així com caracteritzar-ne el comportament i les limitacions.

Metodologia: Estudi descriptiu que inclou 140 referències d'articles de revista extretes de 20 treballs de fi de grau d'Infermeria de la Universitat Complutense de Madrid (2019-2020). Es va elaborar una llista de referències de referència (gold standard) mitjançant revisió per parells. Es va demanar a ChatGPT-4 Turbo que corregís i formatés les referències originals utilitzant una instrucció (prompt) bàsica. Les seves respostes es van comparar amb la llista de referència mitjançant un sistema de codificació de sis categories i el recompte d'accions per element bibliogràfic (n = 521).

Resultats: En conjunt, el 49,3 % de les referències van rebre almenys una intervenció, mentre que el 50,7 % van romandre sense canvis; entre aquestes darreres, 12 ja eren correctes i 59 continuaven contenint errors. De les 521 accions registrades, 130 (25,0 %) van ser adequades i 391 (75,0 %) inadequades (no intervenció: 79,3 %; correcció errònia: 13,0 %; error de format: 7,2 %; al·lucinació: 0,5 %).

Conclusions: Amb una instrucció (prompt) bàsica, ChatGPT-4 Turbo proporciona millores limitades i predomina la no intervenció. La seva aplicació podria ser útil dins d'un protocol de treball (instruccions específiques per a la tasca, verificació de metadades basada en DOI o PMID i supervisió humana), especialment en les fases inicials o com a suport previ a la revisió experta.

Paraules clau

Intel·ligència artificial; avaluació; referències bibliogràfiques; estil Vancouver.

1. Introducción

En la práctica cotidiana de la publicación científica persisten numerosos problemas relacionados con la exactitud de las referencias bibliográficas: errores tipográficos, datos incompletos, omisión de autores, confusiones entre número y volumen o paginaciones imposibles de localizar. Diversos estudios (Montenegro et al. 2021; Armstrong et al. 2018) han revelado que una proporción relevante de las listas de referencias contiene al menos un error que dificulta o incluso impide la identificación de la fuente original, con implicaciones directas para la reproducibilidad de los hallazgos y para la evaluación de la calidad editorial de las revistas.

La correcta elaboración de referencias bibliográficas es, por tanto, un aspecto fundamental en la comunicación científica (Fernández del Castillo et al. 2003), ya que garantiza la trazabilidad de las fuentes, facilita la verificación de los datos y contribuye a la calidad y credibilidad de los trabajos académicos y científicos (Foreman y Kirchhoff 1987; Juncosa Font 1991; Navarro 1998; Todd y Warner 1993). No obstante, la normalización de las referencias ("Style of references standardized" 1970) conforme a estilos de citación específicos puede representar un proceso complejo para autores, estudiantes e incluso editores, debido a la diversidad de estilos bibliográficos existentes (Spence 1980) y a la posibilidad de errores en la transcripción, citación o estructuración de la información.

En ciencias de la salud este problema se atenúa en cierta medida, dado que el estilo más habitual para la redacción de referencias bibliográficas es el conocido como estilo Vancouver (Primo Peña 1996; Patrias y Wendling 2007). Ello permite que los profesionales del ámbito sanitario no tengan que familiarizarse con múltiples estilos, aunque sí resulta imprescindible que conozcan en profundidad la estructura de este estilo, el número de autores a consignar, la forma apropiada de transcribir el nombre de la revista y las diferentes cláusulas (mención del formato en línea, fecha de consulta y disponibilidad de acceso a través de la URL).

Independientemente del estilo utilizado, la redacción y revisión manual de las referencias bibliográficas suponen una carga de trabajo considerable, especialmente en las etapas finales de la elaboración de los manuscritos o durante el proceso de revisión por parte del equipo editorial (González de Zárate Apiñániz et al. 1990). Estas tareas, a menudo percibidas como tediosas, pueden provocar retrasos en la publicación, generar

un elevado número de observaciones por parte de los revisores (*peer review*) e incluso influir negativamente en la aceptación de un manuscrito cuando no se cumplen los requisitos formales solicitados por la revista, como redactar las referencias conforme al estilo requerido o estructurarlas de forma adecuada.

En este escenario han cobrado especial relevancia las herramientas digitales que prometen aliviar parte de esta carga: gestores bibliográficos, complementos integrados en procesadores de texto y, más recientemente, modelos de Inteligencia Artificial (IA) generativa capaces de reescribir o adaptar referencias a un estilo concreto. Estas herramientas podrían agilizar la fase final de la edición y permitir al personal investigador concentrar sus esfuerzos en la interpretación de resultados y en la redacción de la discusión (Coar y Sewell 2010; Nicoll et al. 1996). Sin embargo, su adopción responsable exige conocer no solo su potencial, sino también sus limitaciones y riesgos, especialmente cuando pueden introducir cambios que pasan desapercibidos para el usuario o generar información ficticia, difícil de detectar sin una revisión experta.

La literatura científica reciente ya ha comenzado a explorar el impacto de herramientas específicas de IA en diversas etapas del ciclo de investigación. Se han analizado desde aplicaciones integradas en bases de datos académicas, como Scopus AI, para acelerar el descubrimiento de información y la síntesis de evidencia (Aguilera-Cora et al. 2024), hasta el uso de modelos conversacionales como soporte metodológico en revisiones sistemáticas y *scoping reviews* bajo marcos de trabajo estructurados como SALSA (Lopezosa et al. 2023). Asimismo, autores como Torres-Salinas y Arroyo-Machado (2025) han subrayado la necesidad de una alfabetización crítica en IA, advirtiendo sobre los retos éticos y el "sedentarismo intelectual" que su integración masiva podría suponer para las funciones esenciales de la universidad: la docencia y la investigación académica.

En este contexto, el uso de herramientas basadas en IA, como ChatGPT (Walters y Wilder 2023), ofrece nuevas posibilidades para la automatización, corrección y mejora de la redacción de referencias bibliográficas. Este modelo de lenguaje de gran escala (LLM), desarrollado por OpenAI y basado en la arquitectura *Generative Pre-trained Transformer* (GPT), ha sido entrenado con vastos volúmenes de datos que le confieren una capacidad excepcional para procesar y generar lenguaje natural. Su versatilidad para tareas de revisión y reformateo, sumada a su amplia implantación y accesibilidad mediante una interfaz que no requiere conocimientos técnicos, justifican su selección como objeto de estudio en el presente análisis.

Paralelamente, la creciente integración de estas tecnologías en la comunicación científica ha impulsado el desarrollo de marcos y protocolos destinados a orientar su aplicación responsable. Entre ellos destacan propuestas como MASIA (Marco de Análisis de Sistemas de Inteligencia Artificial) y VERITAS, que ofrecen criterios estructurados para evaluar la idoneidad, fiabilidad y transparencia de la IA en diversos contextos académicos (Codina 2025; Codina et al. 2025). La existencia de dichos marcos pone de manifiesto la necesidad de disponer de evidencia empírica sobre el rendimiento de los modelos en tareas específicas de curación documental, validando la pertinencia de someter a examen su precisión antes de normalizar su uso en los flujos de trabajo editoriales.

No obstante, pese a este marco regulatorio en desarrollo, estas aplicaciones todavía presentan diversas limitaciones (Athaluri et al. 2023). Por consiguiente, antes de recomendar su uso generalizado debería comprobarse empíricamente su aplicabilidad y fiabilidad. Al revisar la literatura científica publicada sobre este tema, no se han identificado trabajos centrados específicamente en las capacidades y limitaciones de las herramientas de IA en la corrección de referencias bibliográficas reales. Hasta ahora, la mayor parte de la evidencia reciente se focaliza en la evaluación de distintos modelos de IA en cuanto a su capacidad para la generación automática de referencias (Sebo 2024; Chelli et al. 2024), con énfasis en las alucinaciones y en los DOIs fabricados, y no en la corrección o normalización de referencias ya existentes. Esta asimetría metodológica

limita la transferencia de los hallazgos a contextos editoriales, donde el reto habitual es estandarizar bibliografías previamente elaboradas.

El presente estudio aborda esta laguna. Su objetivo general es evaluar la utilidad y precisión de ChatGPT-4 Turbo como herramienta para la corrección de referencias bibliográficas de artículos de revista reales, siguiendo el estilo de citación Vancouver. Para ello se plantean tres objetivos específicos: (1) cuantificar la tasa de acciones correctas e incorrectas realizadas por el modelo sobre las referencias fuente; (2) identificar los tipos de errores más frecuentes en las intervenciones del modelo, mediante un sistema de codificación preestablecido; y (3) analizar el comportamiento diferencial del modelo en función del origen geográfico de las publicaciones (internacional, español y latinoamericano).

2. Metodología

Se realizó un estudio descriptivo sobre la calidad de las referencias bibliográficas en una muestra aleatoria de 20 Trabajos de Fin de Grado (TFG) de Enfermería de la Universidad Complutense de Madrid (curso 2019-2020). Partiendo de un universo de 203 trabajos, la selección de esta muestra responde a criterios de viabilidad práctica: el análisis requería la elaboración manual de referencias patrón y una comparación triádica exhaustiva por tres investigadores. Con esta muestra se obtuvieron 140 referencias válidas, consideradas suficientes para un estudio descriptivo y exploratorio orientado a identificar patrones de comportamiento del modelo, y no a producir estimaciones estadísticas poblacionales.

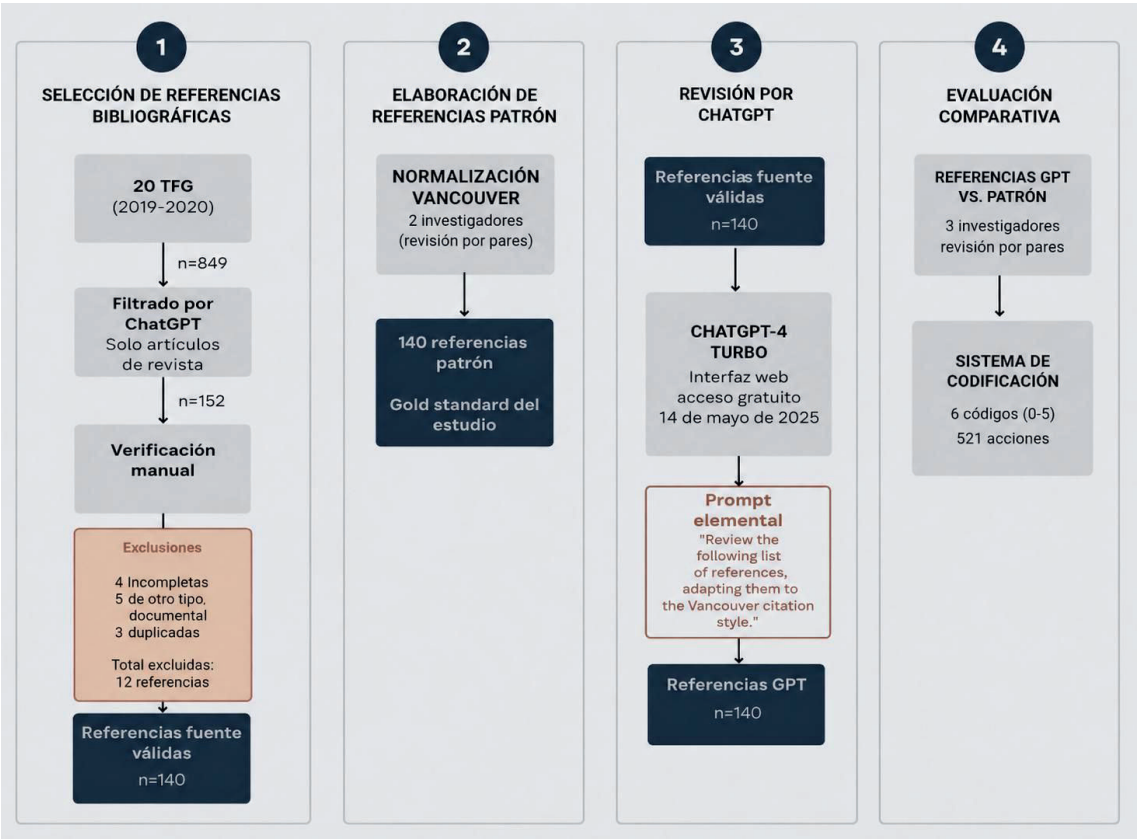

Figura 1. Diagrama de flujo del proceso de selección y análisis de referencias bibliográficas

La elección de los TFG como fuente de información se justifica porque, en etapas iniciales de la formación investigadora del estudiantado, la media de errores por referencia es más elevada (Llopis Agelán et al. 2021), lo que permite evaluar con mayor precisión las capacidades del modelo para detectar y corregir errores. Aunque el estudio se sitúa en el ámbito de las ciencias de la salud (Awrey et al. 2011; Sohn 1983; Taylor 1998), los problemas de normalización de la bibliografía son comunes a otras disciplinas, lo que facilita la extrapolación de las conclusiones y hallazgos a otros campos de la investigación no estrictamente biomédica.

El estudio se desarrolló en cuatro fases (Figura 1):

1. Selección de las referencias bibliográficas. Del conjunto de TFG seleccionados (n = 20) se obtuvieron un total de 849 referencias. La mayoría de estas correspondían a tipos documentales distintos al artículo de revista científica (libros, capítulos de libro, páginas web, guías clínicas, informes institucionales, etc.). Con el objetivo de centrar el análisis en el tipo de referencia con mayor estandarización formal, se solicitó a ChatGPT-4 Turbo que identificara y extrajera exclusivamente las referencias correspondientes a artículos de revista, obteniéndose un total de 152. Posteriormente, uno de los investigadores verificó de forma manual la concordancia entre las referencias extraídas por la herramienta y las presentes en los TFG originales, confirmando su adscripción al tipo documental artículo de revista. Se excluyeron 4 referencias por insuficiencia de datos para su identificación, 5 por corresponder a otros tipos documentales y 3 por estar duplicadas. El resultado final fue de 140 referencias válidas para el análisis (referencias fuente).
2. Normalización de las referencias. Las 140 referencias fuente se normalizaron según el estilo de citación Vancouver mediante revisión por pares (dos investigadores de este estudio) para asegurar precisión, coherencia interna y cumplimiento estricto de la normativa (referencias patrón).
3. Revisión de las referencias bibliográficas. Se solicitó a ChatGPT-4 Turbo la corrección y normalización de las 140 referencias fuente. La consulta se realizó el 14 de mayo de 2025 mediante la interfaz web de ChatGPT-4 Turbo en su modalidad de acceso gratuito, introduciendo las referencias en texto plano para evitar errores de interpretación y sin autenticación, con el fin de limitar los posibles sesgos asociados al historial del usuario. Se utilizó el siguiente *prompt*: "*Review the following list of references, adapting them to the Vancouver citation style*". Se optó deliberadamente por un *prompt* elemental con el propósito de valorar la utilidad real del modelo en contextos no especializados. De este modo, se buscaba simular el comportamiento de usuarios que no poseen formación específica en ingeniería de *prompts* (Torres-Salinas 2025) ni conocimientos avanzados para diseñar consultas complejas o estructuradas (Codina 2024), permitiendo así evaluar el desempeño de la herramienta en su nivel de accesibilidad más básico.
4. Evaluación de las referencias bibliográficas revisadas por ChatGPT-4 Turbo. Se realizó una comparación por pares de las referencias GPT con las referencias patrón, llevada a cabo por tres investigadores de este estudio, mediante un sistema de codificación que clasifica la respuesta del modelo en seis categorías (códigos 0–5). Las primeras cuatro categorías (códigos 0–3) se consideraron resultados erróneos, mientras que las dos últimas categorías (códigos 4–5) se interpretaron como resultados acertados. Así, cada código se definió del siguiente modo:

    - Código 0 (ausencia de corrección): el modelo no realiza ninguna modificación y mantiene elementos vacíos, incompletos o incorrectos presentes en la referencia original.
    - Código 1 (error de formato): el modelo introduce errores de forma, como capitalizar de manera inadecuada los títulos de los artículos o alterar el orden correcto de los distintos elementos de la referencia.

- Código 2 (alucinación): el modelo inventa datos inexistentes en el original, añade información espuria o modifica la existente generando una referencia incorrecta.
- Código 3 (corrección errónea): el modelo rectifica el original alterando los datos y convirtiéndolos en erróneos o eliminando información pertinente y válida.
- Código 4 (mejora): el modelo corrige o completa adecuadamente los datos incorrectos o incompletos de la referencia original y los normaliza según el estilo Vancouver.
- Código 5 (copia fiel): el modelo replica fielmente la referencia que ya estaba correctamente redactada en el original conforme a la normativa.

3. Resultados

De las 140 referencias bibliográficas evaluadas, 69 (49,3 %) recibieron al menos una intervención por parte del modelo de IA, mientras que en 71 (50,7 %) no se registró cambio alguno. De estas últimas, 12 correspondían a referencias correctas en origen que, por tanto, no requerían modificaciones. Sin embargo, las 59 referencias restantes presentaban errores en los que el modelo (ChatGPT-4 Turbo) no intervino, a pesar de que hubiera sido deseable una corrección.

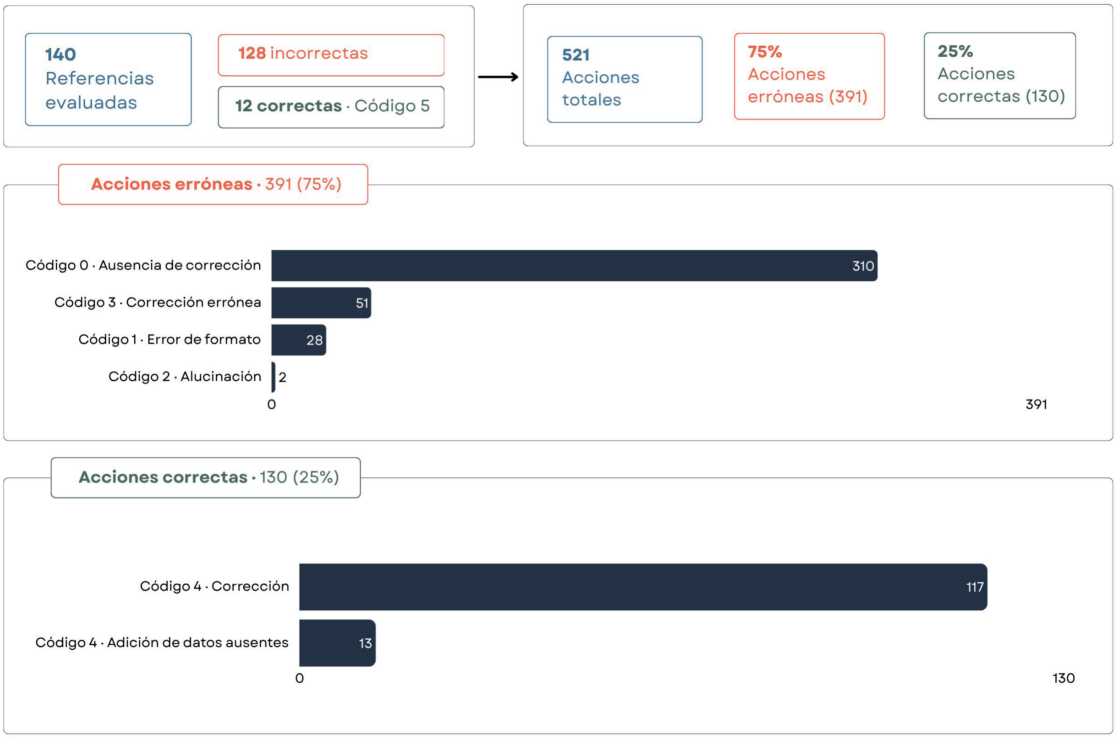

Figura 2. Distribución de las acciones realizadas por ChatGPT-4 Turbo según el sistema de codificación

En el conjunto de las 128 referencias bibliográficas incorrectas en origen (91,4 %), el modelo realizó 521 acciones a nivel de elemento: 130 (25,0 %) fueron calificadas como adecuadas y 391 (75,0 %) como erróneas. Entre las 130 acciones adecuadas, 13 (10,0 %) consistieron en la adición de datos ausentes en las referencias fuente y 117 (90,0 %) en la rectificación de datos que habían sido redactados de forma incorrecta en dichas referencias fuente. Las principales mejoras se distribuyeron entre los distintos elementos bibliográficos del modo siguiente: 25 en el título de la revista (19,2 %), 19 en el elemento fecha (14,6 %) y 50 en la signatura, correspondiendo 32 a la redacción de las páginas (24,6 %) y 18 a la consignación del volumen (13,8 %). Entre las 391 acciones erróneas, 310

se clasificaron como ausencias de corrección (79,3 %), 51 como correcciones erróneas (13,0 %), 28 como errores de formato (7,2 %) y 2 como alucinaciones (0,5 %) (Figura 2).

Respecto al total de acciones realizadas por la herramienta (n = 521) en cada uno de los elementos que conforman una referencia bibliográfica, se observó que, entre las de carácter general, destacaban los errores ortográficos, mecanográficos y de estilo. Entre estos se incluyeron 56 errores en la capitalización de títulos (10,8 %), 34 inclusiones de URL incorrectas (6,5 %) y 10 adiciones de elementos innecesarios (1,9 %). En el elemento de la Autoría se registraron 40 modificaciones erróneas relacionadas con abreviaturas o escrituras incorrectas de los nombres y apellidos de los autores (7,7 %), frente a tan solo 7 mejoras (1,3 %) o correcciones adecuadas. Asimismo, en el uso de la cláusula "et al." se identificaron 9 correcciones erróneas (1,7 %) (Tabla 1).

Tabla 1. Correcciones en referencias por elemento bibliográfico

Elemento	Subtipo de error	N° modif. correctas (cód. 4-5)	N° modif. incorrectas (cód. 0-3)	% correctas	% incorrectas
General	Se añaden elementos innecesarios	0	10	0,0%	1,9%
	Errores mecanográficos/ ortográficos	11	56	2,1%	10,8%
	URL incorrecta	4	34	0,8%	6,5%
Autoría	Ausencia de campo	0	4	0,0%	0,8%
	Orden incorrecto de elementos	1	0	0,2%	0,0%
	No abrevia o incluye errores	7	40	1,3%	7,7%
	Puntuación incorrecta	0	0	0,0%	0,0%
	Uso incorrecto de "et al."	0	9	0,0%	1,7%
	Lugar equivocado	0	0	0,0%	0,0%
Título	Ausencia del campo	0	0	0,0%	0,0%
	Título incompleto o erróneo	7	10	1,3%	1,9%
Revista	Ausencia del campo	4	1	0,8%	0,2%
	Fallo en abreviatura o cursiva	21	51	4,0%	9,8%
Fecha	Ausencia del campo	0	0	0,0%	0,0%
	Lugar equivocado	1	0	0,2%	0,0%
	Puntuación incorrecta	18	19	3,5%	3,7%
	Datos incorrectos	0	5	0,0%	1,0%
Volumen	Ausencia del campo	4	4	0,8%	0,8%
	Puntuación incorrecta	12	44	2,3%	8,5%
	Lugar equivocado	2	0	0,4%	0,0%
Número	Ausencia del campo	3	19	0,6%	3,7%
	Lugar equivocado	1	0	0,2%	0,0%
	Puntuación incorrecta	2	33	0,4%	6,3%
Páginas	Ausencia del campo	2	11	0,4%	2,1%
	Errores de paginación o formato	12	41	2,3%	7,9%
	No se abrevian	18	0	3,5%	0,0%
	Lugar equivocado	0	0	0,0%	0,0%
TOTAL		130	391	25,0%	75,0%

Todas las referencias analizadas incluían el elemento Título y se obtuvieron 10 correcciones erróneas (1,9 %) frente a 7 (1,3 %) mejoras relativas a títulos incompletos o erróneos. En relación con el título de la fuente o publicación se detectaron 51 modificaciones incorrectas (9,8 %), fundamentalmente asociadas a errores en las abreviaturas de las revistas o al uso de la cursiva (no requerida por el estilo Vancouver, aunque sí habitual en otros estilos), frente a 21 mejoras (4,0 %) que incorporaron una corrección adecuada de la abreviatura. En lo que respecta a la signatura, las equivocaciones afectaron

principalmente a cuestiones de puntuación —19 en la fecha (3,7 %), 44 en el volumen (8,5 %) y 33 en el número (6,3 %)— y a la paginación, con 41 errores en la descripción de las páginas (7,9 %).

Por último, se observó que, en cuanto a la procedencia geográfica de las publicaciones, la media de errores por referencia fue menor en las revistas internacionales (2,3 errores por referencia), mientras que resultó más elevada en las revistas españolas (3,9) y latinoamericanas (3,8).

4. Discusión

El reducido número de referencias correctas (n = 12) en el conjunto de la bibliografía redactada por los alumnos en sus TFG (8,6 %) sugiere, por un lado, una posible falta de formación específica y, por otro, cierto desinterés por la correcta elaboración de la bibliografía, que a menudo se percibe como una fase de menor importancia frente a otros apartados de los trabajos académicos, como la hipótesis, la metodología, los objetivos o la discusión. Este hallazgo es coherente con los resultados de Llopis Agelán et al. (2021), que ya identificaron medias de errores por referencia elevadas en TFG de Enfermería, y subraya la necesidad de reforzar la formación de los estudiantes en competencias informacionales, ya que una correcta normalización de la bibliografía refleja un uso adecuado de las fuentes, facilita la evaluación de cada TFG por parte del profesorado y contribuye a que futuros lectores puedan identificar, localizar y recuperar los documentos consultados.

En respuesta al primer objetivo, los resultados muestran que, sobre las 128 referencias incorrectas en origen, ChatGPT-4 Turbo realizó 521 acciones, de las cuales solo el 25,0 % fueron clasificadas como adecuadas. Este desempeño limitado es consistente con los hallazgos de Walters y Wilder (2023) y de Chelli et al. (2024), que reportaron tasas de error relevantes en tareas de generación bibliográfica automatizada; sin embargo, el presente estudio aporta evidencia específica sobre un escenario metodológicamente distinto y más complejo: la corrección de referencias ya existentes. A diferencia de la generación —que parte de información estructurada para producir una referencia nueva—, la corrección exige que el modelo reconozca el estilo empleado, lo compare implícitamente con una norma patrón y proponga una nueva construcción elemento por elemento (autores, título, revista, signatura), tomando decisiones sobre la pertinencia o la exactitud de cada dato. Este segundo escenario resulta intrínsecamente más exigente, y los resultados sugieren que los modelos actuales no están aún suficientemente optimizados para abordarlo con garantías en contextos no especializados.

Por otra parte, se observó una elevada tasa de referencias en las que ChatGPT-4 Turbo no realizó modificación alguna, siendo esta "no intervención" el comportamiento dominante en el estudio. Aunque este hecho podría desincentivar el uso de la IA para estas tareas, resulta pertinente analizar si esta actitud pasiva se relaciona nuevamente con el uso de un *prompt* poco exigente o con la dificultad del modelo para identificar correctamente las referencias en Internet y realizar las correcciones oportunas. Cuando la referencia bibliográfica está redactada de forma razonablemente correcta, a la IA le resulta más sencillo localizarla en recursos accesibles como PubMed o Google Scholar. En cambio, si la referencia presenta errores importantes desde el inicio —por ejemplo, en la autoría—, la identificación del artículo original se complica considerablemente.

En cuanto al segundo objetivo, el análisis de los tipos de error revela que las ausencias de corrección (código 0) representaron el 79,3 % de las acciones erróneas (310 de 391), seguidas de las correcciones erróneas (código 3; 13,0 %), los errores de formato (código 1; 7,2 %) y, de forma residual, las alucinaciones (código 2; 0,5 %). Este último dato merece

una atención especial: frente a la percepción generalizada de que la IA genera numerosas alucinaciones, en el presente estudio se observó una tasa muy baja de este tipo de actuaciones. Este hallazgo contrasta con otros trabajos centrados en la generación de contenido (Athaluri et al. 2023; Sebo 2024), lo que sugiere que, en tareas fuertemente normalizadas como la edición de referencias, el modelo tiende más a la inacción que a la invención. Si los modelos de IA presentan bajas tasas de invención en tareas de corrección, pueden ofrecer mayor seguridad a los revisores editoriales a la hora de emplearlos como apoyo en la revisión de la bibliografía, siempre que exista supervisión humana.

Una de las principales contribuciones positivas identificadas se relaciona con los aspectos formales de las referencias, especialmente la puntuación en los elementos de volumen y número o las abreviaturas en la paginación. Este tipo de correcciones refuerza la utilidad de la IA cuando se parte de información relativamente normalizada, ya que, conociendo qué información debe consignarse en cada elemento y cómo se relacionan las distintas partes de la referencia (por ejemplo, el uso del punto y coma tras el año, los paréntesis para el número y los dos puntos antes de la paginación), los modelos tienen más facilidad para realizar ajustes mecánicos. No obstante, en su configuración actual, esta capacidad no parece aportar una ventaja sustancial frente a herramientas tradicionales como los gestores bibliográficos, que realizan estas tareas de forma precisa y automatizada siempre que se les proporcione una información bibliográfica completa y correcta.

Precisamente aquí reside uno de los principales retos: cuando un usuario alimenta su propia base de datos en un gestor bibliográfico conoce el origen de la información (PubMed, la web de la revista, etc.), mientras que el funcionamiento poco transparente de muchos sistemas de IA no permite conocer con exactitud qué fuentes consulta ni cómo integran la información, que puede aparecer redactada de forma muy heterogénea en la red. Basta examinar cuántas variantes de una misma referencia pueden localizarse en Google Scholar o ResearchGate para constatar la dificultad de elegir la versión más adecuada.

Más allá de las correcciones centradas en el formato, uno de los aportes más prometedores de la IA a la corrección de referencias es la incorporación o enriquecimiento de datos que no fueron consignados en la referencia fuente (por ejemplo, la inclusión de la URL o del DOI), mejorando así el comportamiento puramente mecánico de los gestores bibliográficos. Si un usuario introduce en un gestor un conjunto de referencias incompletas, la bibliografía obtenida será inevitablemente incompleta. En cambio, los modelos actuales de IA ya son capaces, aunque todavía en un porcentaje limitado, de completar la información consultando fuentes adecuadas y extrayendo de ellas los metadatos necesarios. En este sentido, no resulta descabellado plantear que, en el futuro, los gestores bibliográficos integren capacidades de IA que permitan enriquecer de manera notable las referencias aportadas por los usuarios a partir de la información disponible en bases de datos o en la propia web de la revista.

Respecto al tercer objetivo, es significativo que el modelo obtuviera mejores resultados en la corrección de referencias procedentes de revistas publicadas en el ámbito anglosajón que en el ámbito hispanohablante. Esta diferencia puede deberse al factor idioma (Gonzalez Fiol et al. 2025), a la mayor accesibilidad de las fuentes internacionales (como PubMed) o a la dificultad de obtener los metadatos necesarios para corregir o completar las referencias en revistas españolas y latinoamericanas. Resulta llamativo que persistan errores incluso cuando existe un recurso de alta visibilidad como SciELO para las publicaciones hispanohablantes, lo que sugiere que el modelo no lo prioriza en ausencia de instrucciones explícitas al respecto.

5. Conclusiones

Los resultados de este estudio permiten concluir que ChatGPT-4 Turbo, empleado con un *prompt* elemental, presenta una utilidad limitada para la corrección de referencias bibliográficas en estilo Vancouver. Solo el 25,0 % de las acciones registradas sobre referencias incorrectas fueron clasificadas como adecuadas, y el comportamiento dominante del modelo fue la "no intervención", con independencia de la presencia o ausencia de errores en la referencia fuente. Las alucinaciones fueron marginales (0,5 %), lo que contrasta con estudios previos centrados en la generación bibliográfica y sugiere que las tareas de corrección activan un patrón de cautela diferente al de la generación. Las mejoras más frecuentes se concentraron en aspectos formales de la signatura —paginación, volumen, abreviatura de revista—, mientras que los errores de autoría y los problemas de localización de fuentes hispanohablantes resultaron más resistentes a la corrección automática.

Es importante considerar que el tiempo requerido para realizar una revisión mediante estas herramientas es significativamente menor que el de los métodos tradicionales, como la captura manual o la corrección pormenorizada de metadatos en gestores bibliográficos. Esta relación favorable entre esfuerzo y resultado posiciona a la IA como un recurso valioso en tareas de apoyo, siempre que medie una supervisión experta. Bajo esta premisa, se recomienda que los docentes incorporen advertencias explícitas sobre la fiabilidad de la IA en las guías académicas, subrayando la necesidad de una verificación final humana. Por su parte, los equipos editoriales deberían supeditar el uso de estas herramientas a protocolos definidos y *prompts* estructurados, planteando como horizonte una futura integración de la IA en los sistemas de gestión editorial que permita automatizar la detección de errores.

De cara al desarrollo futuro de esta línea de investigación, se identifican tres vías prioritarias. En primer lugar, el diseño y evaluación de *prompts* de mayor complejidad que orienten al modelo a trabajar elemento a elemento, especificando fuentes de consulta prioritarias (PubMed, DOI, SciELO para publicaciones hispanohablantes) e instruyéndolo para que informe explícitamente sobre si ha podido verificar cada referencia en una base de datos externa. Un *prompt* estructurado de estas características podría reducir significativamente la tasa de "no intervención" y mejorar la transparencia del proceso. En segundo lugar, la integración de capacidades de IA en procesadores de texto y gestores bibliográficos existentes representa un horizonte especialmente prometedor: a diferencia del uso aislado de un modelo de lenguaje, un gestor con IA incorporada podría enriquecer de forma automatizada las referencias incompletas a partir de los metadatos disponibles en bases de datos, combinando la precisión normativa del gestor con la capacidad inferencial del modelo. En tercer lugar, la replicación de este estudio con modelos más recientes y muestras de referencias de otras disciplinas y estilos de citación permitiría establecer hasta qué punto los resultados aquí obtenidos son generalizables o están condicionados por el contexto específico del estudio.

6. Referencias

Aguilera-Cora, Elisenda, Carlos Lopezosa, José Fernández-Cavia, y Lluís Codina. 2024. "Accelerating Research Processes with Scopus AI: A Place Branding Case Study". *Revista Panamericana de Comunicación* 6 (1). https://doi.org/10.21555/rpc.v6i1.3088.

Armstrong, Michael F., Joseph H. Conduff, John E. Fenton, y Daniel H. Coelho. 2018. "Reference Errors in Otolaryngology-Head and Neck Surgery Literature". *Otolaryngology - Head and Neck Surgery* 159 (2): 249-53. https://doi.org/10.1177/0194599818772521.

Athaluri, Sai Anirudh, Sandeep Varma Manthena, V. S. R. Krishna Manoj Kesapragada, Vineel Yarlagadda, Tirth Dave, y Rama Tulasi Siri Duddumpudi. 2023. "Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence Hallucination in Scientific Writing through ChatGPT References". *Cureus* 15 (4): e37432. https://doi.org/10.7759/cureus.37432.

Awrey, Julianne, Kenji Inaba, Galinos Barmparas, et al. 2011. "Reference Accuracy in the General Surgery Literature". *World Journal of Surgery* 35 (3): 475-79. https://doi.org/10.1007/s00268-010-0912-7.

Chelli, Mikaël, Jules Descamps, Vincent Lavoué, et al. 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.

Coar, Jaekea T., y Jeanne P. Sewell. 2010. "Zotero: Harnessing the Power of a Personal Bibliographic Manager". *Nurse Educator* 35 (5): 205-7. https://doi.org/10.1097/NNE.0b013e3181ed81e4.

Codina, Lluís. 2024. "12 prompts para usar la inteligencia artificial en la academia". *Lluís Codina*, abril 8. https://www.lluiscodina.com/prompts-inteligencia-artificial-academia/.

Codina, Lluís. 2025. "Uso ético y eficiente de la inteligencia artificial en trabajos académicos: Veritas e interacción crítica escalonada". *BiD: textos universitaris de biblioteconomia i documentació*, n.º 55: 1-9. https://doi.org/10.1344/bid2025.55.01.

Codina, Lluís, Elisenda Aguilera-Cora, Carlos Lopezosa, y Pere Freixa Font. 2025. "Pensamiento crítico e inteligencia artificial en la academia: procedimiento de análisis matricial cualitativo para evaluar sistemas de IA". En Comunicación digital: tendencias y buenas prácticas. Guallar J, Vállez M, Ventura-Cisquella A, coordinadores. Ediciones Profesionales de la Información, p.171–183. https://doi.org/10.3145/cuvicom.12.esp.

Fernández del Castillo, Carlos, Jorge Delgado Urdapilleta, y Ernesto Castelazo Morales. 2003. "¿Por qué son importantes las referencias bibliográficas?" *Ginecologia y Obstetricia de Mexico* 71: 275-76.

Foreman, M. D., y K. T. Kirchhoff. 1987. "Accuracy of References in Nursing Journals". *Research in Nursing & Health* 10 (3): 177-83. https://doi.org/10.1002/nur.4770100310.

González de Zárate Apiñániz, P., MA García Martín, y C. Aguirre Errasti. 1990. "Búsqueda, selección crítica y cita de referencias bibliográficas en los artículos médicos". *Medicina Clínica (Barcelona)* 95 (1): 39.

Gonzalez Fiol, Antonio, Allison A. Mootz, Zili He, Carlos Delgado, Vilma Ortiz, y Sharon C. Reale. 2025. "Accuracy of Spanish and English-generated ChatGPT responses to commonly asked patient questions about labor epidurals: a survey-based study among bilingual obstetric anesthesia experts". *International Journal of Obstetric Anesthesia* 61: 104290. https://doi.org/10.1016/j.ijoa.2024.104290.

Juncosa Font, S. 1991. "Incorrecciones en las citas bibliográficas: un obstáculo para la calidad de nuestras publicaciones científicas". *Atención Primaria* 8 (6): 515-16.

Llopis Agelán, José Luis, Óliver Martín Martín, José Manuel Estrada Lorenzo, y Ramón del Gallego Lastra. 2021. "Precisión de las referencias bibliográficas de los Trabajos de Fin de Grado en Enfermería". *Metas de enfermería* 24 (6): 5-10. https://doi.org/10.35667/METASENF.2021.24.1003081783.

Lopezosa, Carlos, Lluís Codina, y Núria Ferran Ferrer. 2023. *ChatGPT como apoyo a las systematic scoping reviews: integrando la inteligencia artificial con el framework SALSA*. Universitat de Barcelona. https://hdl.handle.net/2445/193691.

Montenegro, Thiago S., Kevin Hines, Paul P. Partyka, y James Harrop. 2021. "Reference Accuracy in Spine Surgery". *Journal of Neurosurgery. Spine* 34 (1): 22-26. https://doi.org/10.3171/2020.6.SPINE20640.

Navarro, FA. 1998. "Inexactitud de las referencias bibliográficas: ¿se citan artículos no leídos?" *Revista Clínica Española* 198 (2): 117.

Nicoll, L. H., T. H. Ouellette, D. C. Bird, J. Harper, y J. Kelley. 1996. "Bibliography Database Managers. A Comparative Review". *Computers in Nursing* 14 (1): 45-56.

Patrias, Karen, y Dan Wendling. 2007. *Citing medicine*. 2nd ed. National Library of Medicine (US).

Primo Peña, Elena. 1996. "El estilo Vancouver". Jano: Medicina y Humanidades. *Jano: Medicina y humanidades* 50 (1151): 76-77. https://dialnet.unirioja.es/servlet/articulo?codigo=9687912.

Sebo, Paul. 2024. "How Accurate Are the References Generated by ChatGPT in Internal Medicine?" *Internal and Emergency Medicine* 19 (1): 247-49. https://doi.org/10.1007/s11739-023-03484-5.

Sohn, CA. 1983. "Errors in commonly used references". *American Journal of Hospital Pharmacy* 40 (10): 1622.

Spence, AA. 1980. "Not all the way on the vancouver style". *British Journal of Anaesthesia* 52 (5): 469-70. https://academic.oup.com/bja/article/52/5/469/319217.

"Style of references standardized". 1970. *New England Journal of Medicine* 282 (1): 49-49. https://doi.org/10.1056/NEJM197001012820115.

Taylor, Mary K. 1998. "The Practical Effects of Errors in Reference Lists in Nursing Research Journals". *Nursing Research* 47 (5): 300-303. https://doi.org/10.1097/00006199-199809000-00010.

Todd, Michael M., y David S. Warner. 1993. "On the Importance of Inaccurate Bibliographic Citations". *Anesthesiology* 78 (3): 615-16. https://doi.org/10.1097/00000542-199303000-00043.

Torres-Salinas, Daniel. 2025. *Primeros pasos con ChatGPT e ingeniería de prompts*. Universidad de Granada. https://doi.org/10.5281/zenodo.15260567.

Torres-Salinas, Daniel, y Wenceslao Arroyo-Machado. 2025. "Enseñar e investigar con inteligencia artificial: una llamada a la reflexión – BID". Sec. 1-8. *BiD: textos universitaris de biblioteconomia i documentació* 54 (2). https://doi.org/10.1344/BID2025.54.01.

Walters, William H., y Esther Isabelle Wilder. 2023. "Fabrication and Errors in the Bibliographic Citations Generated by ChatGPT". *Scientific Reports* 13 (1): 14045. https://doi.org/10.1038/s41598-023-41032-5.