ChatGPT podría ser el revisor de tu próximo artículo científico. Evidencias de los límites de las revisiones académicas asistidas por inteligencia artificial

ChatGPT could be the reviewer of your next scientific paper. Evidence on the limits of AI-assisted academic reviews

David Carabantes; José L. González-Geraldo; Gonzalo Jover

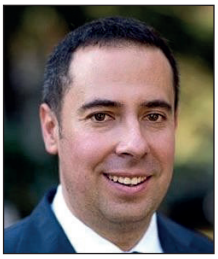
Note: This article can be read in its English original version on:
https://revista.profesionaldelainformacion.com/index.php/EPI/article/view/87376

Cómo citar este artículo.
Este artículo es una traducción. Por favor cite el original inglés:

Carabantes, David; **González-Geraldo, José L.**; **Jover, Gonzalo** (2023). "*ChatGPT* could be the reviewer of your next scientific paper. Evidence on the limits of AI-assisted academic reviews". *Profesional de la información*, v. 32, n. 5, e320516.

https://doi.org/10.3145/epi.2023.sep.16

Artículo recibido el 22-05-2023
Aceptación definitiva: 29-08-2023



David Carabantes
https://orcid.org/0000-0001-9897-4847

Universidad Complutense de Madrid
Facultad de Educación
Rector Royo Villanova, 1
28040 Madrid, España
dcaraban@ucm.es

José L. González-Geraldo ✉
https://orcid.org/0000-0003-1698-0122

Facultad de Ciencias de la Educación y Humanidades
Avda. Los Alfares, 44
16002 Cuenca, España
joseluis.ggeraldo@uclm.es

Gonzalo Jover
https://orcid.org/0000-0002-6373-4111

Universidad Complutense de Madrid
Facultad de Educación
Rector Royo Villanova, 1
28040 Madrid, España
gjover@ucm.es

Resumen

La irrupción de la inteligencia artificial (IA) en todos los ámbitos de nuestra vida es una realidad a la que la universidad, como institución de educación superior, ha de responder con prudencia, pero también con decisión. El presente artículo discute el potencial que recursos basados en la IA presentan como potenciales evaluadores de artículos científicos en una hipotética revisión por pares de artículos ya publicados. A través de distintos modelos (*GPT-3.5* y *GPT-4*) y plataformas (*ChatPDF* y *Bing*), obtuvimos tres revisiones completas, tanto cualitativas como cuantitativas, para cada uno de los cinco artículos examinados, pudiendo así delinear y contrastar los resultados de todas ellas en función de las revisiones humanas que estos mismos artículos recibieron en su momento. Las evidencias encontradas ponen de relieve hasta qué punto podemos y debemos confiar en los modelos de lenguaje generativos para sostener nuestras decisiones como expertos cualificados en nuestro campo. Asimismo, los resultados corroboran las alucinaciones propias de estos modelos al mismo tiempo que señalan uno de sus grandes defectos actuales: el límite de la ventana contextual. Por otro lado, el estudio también señala las bondades inherentes de un modelo que se encuentra en plena fase de expansión, propor-

cionando una visión detallada del potencial y las limitaciones que estos modelos ofrecen como posibles asistentes a la revisión de artículos científicos, proceso clave en la comunicación y difusión de la investigación académica.

Palabras clave

Inteligencia artificial; IA; Inteligencia artificial generativa; Ventana contextual; *ChatGPT*; *ChatPDF*; *Bing;* Revisión asistida por IA; Revisión por pares; Revisión académica; Publicación académica; Comunicación científica.

Abstract

The irruption of artificial intelligence (AI) in all areas of our lives is a reality to which the university, as an institution of higher education, must respond prudently, but also with no hesitation. This paper discusses the potential that resources based on AI presents as potential reviewers of scientific articles in a hypothetical peer review of already published articles. Using different models (*GPT-3.5* and *GPT-4*) and platforms (*ChatPDF* and *Bing*), we obtained three full reviews, both qualitative and quantitative, for each of the five articles examined, thus being able to delineate and contrast the results of all of them in terms of the human reviews that these same articles received at the time. The evidence found highlights the extent to which we can and should rely on generative language models to support our decisions as qualified experts in our field. Furthermore, the results also corroborate the hallucinations inherent in these models while pointing out one of their current major shortcomings: the context window limit. On the other hand, the study also points out the inherent benefits of a model that is in a clear expansion phase, providing a detailed view of the potential and limitations that these models offer as possible assistants to the review of scientific articles, a key process in the communication and dissemination of academic research.

Keywords

Artificial intelligence; AI; Generative artificial intelligence; Contextual window; *ChatGPT*; *ChatPDF*; *Bing*; AI-assisted review; Peer review; Academic review; Academic publication; Scientific communication.

Financiación

Este artículo ha sido parcialmente financiado por el *Grupo de Investigación Cultura Cívica y Políticas Educativas*, de la *Universidad Complutense de Madrid*, mediante el *Programa de Financiación de Grupos de Investigación* de la *UCM* (GRFN32/23).

1. Introducción

Como ha sucedido con otras novedades que en su día marcaron época, es probable que en no mucho tiempo el nombre de *ChatGPT*, hoy en boca de todo el mundo, se difumine, y aparezcan otras marcas y logotipos que, recogiendo entre otros avances la herencia del procesamiento de lenguaje natural (NLP: *natural language processing*) o los grandes modelos de lenguaje (LLM: *large language models*), encarnen la transformación GPT (*generative pre-trained transformer*) pre-entrenada, generativa y basada en la revolución de los conocidos como *transformers* (**Vaswani** *et al.*, 2017) de una manera cualitativamente más compleja y fiable (**González-Geraldo**; **Ortega-López**, 2023). Junto a *ChatGPT* (*OpenAI*) hoy *Bing* (*Microsoft*), *Bard* (*Google*) y *Claude* (*Anthropic*) parecen ser las apuestas principales entre estos recursos.

La discusión educativa en torno a la emergencia de este tipo de innovaciones no es nueva, pero en este momento la popularización de *ChatGPT*, como sinónimo de inteligencia artificial (IA) ha generado un debate en el que parece debemos optar entre el pánico o la disrupción (**García-Peñalvo**, 2023) los peligros y las promesas (**Jalil** *et al.*, 2023) los retos y las oportunidades (**Kasneci** *et al.*, 2023) los riesgos y los beneficios (**Sok**; **Heng**, 2023). Se ha llegado a plantear si estamos, por ejemplo, ante el fin de la evaluación tradicional en educación superior (**Rudolph**; **Tan**; **Tan**, 2023) o si, en el fondo, estamos ante un demonio o nuestro ángel de la guarda (**Tlili** *et al.*, 2023). Sea como fuere, el binomio entre educación e IA se ha de poner al servicio, como en otras profesiones, del bien social (**Peña-Fernández** *et al.*, 2023). Aunque las universidades están empezando a regular el uso de estos recursos, como sucede en otros temas que demandan una respuesta rápida, todavía falta aquí una política académica suficientemente coordinada (**Álvarez-Castillo**; **Fernández-Caminero**, 2023).

Como investigadores, creemos oportuno abrazar la realidad de la IA para examinar la intersección que existe entre la inevitabilidad de su advenimiento y la posibilidad de asumir sus potencialidades, al tiempo que minimizamos sus puntos débiles, en especial los que atañen a la ética (**Crawford**; **Cowling**; **Allen**, 2023) y a la integridad académica (**Perkins**, 2023; **Chomsky**; **Roberts**; **Watumull**, 2023). Nuestro objetivo aquí, en este sentido, es analizar las posibilidades de los modelos generativos de textos basados en IA para llevar a cabo la *peer-review* de artículos científicos propuestos para ser publicados.

2. Justificación y estado de la cuestión

Es evidente la repercusión que la *peer-review* tiene en la mejora continua del proceso de publicación científica. La revisión por pares normalmente es ciega, aunque no siempre de manera exclusiva y, según el área de conocimiento, puede actuar junto a otros mecanismos, como la revisión abierta. El procedimiento tiene sus detractores, si bien las críticas y alternativas al mismo se asientan en premisas no siempre compartidas (**Campanario**, 1998a; **Campanario**, 1998b). El uso de la IA ha venido a sumarse a esta discusión.

Entre las limitaciones que se achacan a la revisión por pares están los posibles sesgos personales de los evaluadores, los conflictos de interés, la variabilidad de la calidad e inconsistencias en las revisiones por disparidad en el grado de profundización en el proceso de evaluación, la dificultad de encontrar académicos especializados disponibles, el tiempo de respuesta a la solicitud de participación y el plazo de entrega de la revisión. En este artículo pretendemos determinar si los softwares de IA, del estilo al popular *ChatGPT*, pueden ser una solución efectiva para resolver algunos de estos problemas y retos, según se sugiere ya en algunos entornos especializados, como *Scholarcy* y *Researchleap*:
https://www.scholarcy.com/how-reviewers-can-use-ai-right-now-to-make-peer-review-easier
https://researchleap.com/ai-and-the-future-of-academic-publishing-how-artificial-intelligence-is-transforming-the-peer-review-process

Hay que conocer de antemano cómo funcionan estos recursos sustentados en la arquitectura de modelo de lenguaje *GPT*. Sin entrar en detalles, de manera ciertamente compleja y no siempre bajo control –lo que produce el efecto conocido como la "caja negra" (**Zhai,** 2023)– *GPT* transforma palabras en *tokens*, partes de palabras, para posteriormente reconstruir estos *tokens* en un discurso coherente, dando lugar a lo que **Marcus** (2022) denomina un "pastiche". Aquí radica una de las maravillosas potencialidades que nos ofrece: su *output* es creativo, nunca originado antes, pero siempre basado en lo que existió, con lo que fue entrenado, millones de datos de los diversos ámbitos de conocimiento que le permiten producir textos especializados. De hecho, los propios trabajadores de *OpenAI* no tardaron en avisar de que uno de los malos usos potenciales que estas aplicaciones pueden acarrear es el de la escritura académica fraudulenta (**Brown** *et al.*, 2020, p. 35). Por el lado positivo, su capacidad hace que las mismas puedan convertirse también en un instrumento de la revisión por pares, facilitando una mayor y más eficiente gestión de la sobrecarga que sufren las revistas, sobre todo las que pasan determinados umbrales de ciertos catálogos.

En los últimos tiempos, varios editoriales de revistas científicas han expresado su expectación con respecto al uso de la IA y de recursos como *ChatGPT* (**Švab**; **Klemenc-Ketiš**; **Zupanič**, 2023; **Lira** *et al.*, 2023) y han emergido trabajos académicos que exploran sus principales desafíos, virtudes y limitaciones en los procesos de revisión, dentro de una tendencia general a la automatización de estos procesos (**Checco** *et al.*, 2021; **Severin** *et al.*, 2022; **Srivastava**, 2023). Hay quienes abogan por la necesidad de establecer sin dilación protocolos sobre la utilización de estas herramientas en la revisión por pares (**García**, 2023) sugiriéndose la conveniencia de especificar el uso de IA en las revistas para, por ejemplo, verificar el cumplimiento de las políticas editoriales, resumir contenido o identificar debilidades y fortalezas del manuscrito (**Hosseini**; **Horbach**, 2023).

Nuestra aportación intenta avanzar en esta línea, no solo en el aspecto académico, sino también en algunas limitaciones técnicas que, hoy por hoy, condicionan las posibilidades de estos recursos en la revisión científica de los trabajos. La que más nos interesa es la relativa a la "ventana contextual". Resumidamente, este concepto alude a la extensión de palabras –recordemos, *tokens*– que el modelo puede tener presente a la hora de generar sus resultados. Esta ventana contextual es 4K en la versión *GPT-3.5* (4.096 *tokens*). Dicho de otra forma, el modelo no puede "recordar" más de 3.000 palabras, aproximadamente. De hecho, no recuerda nada, pues no hay memoria de por medio. Cada vez que mandamos una nueva consulta, el sistema contabiliza los *tokens* de nuestro *prompt*, le suma *n tokens* de la conversación anterior hasta cubrir el cupo y toma en consideración el conjunto, siempre bajo el límite de 4K *tokens*. Una limitación que, sin duda, pone en tela de juicio cualquier revisión que pueda realizar, pues raro es el artículo especializado que no llega a las 5.000-6.000 palabras como mínimo.

El avance conseguido con *GPT-4*, lanzado el 14 de marzo de 2023, fue ciertamente esperanzador en este aspecto. La ventana contextual pasó de 4K a 8K, consiguiendo así un significativo aumento a 8.192 *tokens* –por encima de las 6.000 palabras– e incluso mucho más con una versión de 32K –sobre los 32.000 *tokens*–, unas 25.000 palabras, bastante superior de lo necesario para poder revisar artículos e incluso otras producciones mayores. El andamiaje de esta investigación, creímos, estaba asegurado tras el anuncio de *GPT-4*. No obstante, a día de hoy –mayo de 2023–, la ventana contextual sigue siendo la misma usada por su versión anterior (4K) quedando reservadas las dos superiores (8K-32K) para desarrolladores.

Además del límite del contexto, *ChatGPT* ofrece otro límite que se explica por su propia estructura. Al ser una plataforma que emula la conversación natural del ser humano, está diseñada para que las interacciones entre el usuario y el recurso sean más o menos cortas, como en una conversación. Al ser un *chatbot*, es lógico entender que presente también un límite de *tokens* de entrada para cada consulta que le hagamos. En estos momentos, y por razones similares a las ya esgrimidas, tanto las variantes *ChatGPT-3.5* como *ChatGPT-4* no suelen aceptar entradas que superen las 2.200 palabras. Dicho de otro modo, aunque quisiéramos introducir artículos poco a poco para después preguntarle sobre ellos, la realidad sería que solo se quedaría con información parcial, principalmente de los dos últimos *inputs*.

3. Procedimiento

Aun con las limitaciones que hemos señalado, creemos necesario y posible profundizar en las posibilidades que estos recursos nos pueden ofrecer para evaluar artículos. Para ello, acudimos a dos plataformas que, incluso compartiendo las limitaciones de una ventana contextual de 4K, nos han ayudado a solventar el segundo de los problemas (límite por *input*), permitiéndonos facilitarles archivos en formato PDF.

La primera plataforma usada –que denominaremos revisores 1 y 3– fue *ChatPDF* (basada en *GPT-3.5 turbo*), la segunda –revisor 2– el propio navegador *Microsoft Edge*, pues también puede usarse como lector de archivos PDF y actuar

con el asistente *Bing* (basado en *GPT-4*). De esta manera obtenemos tres revisiones independientes que nos permiten contrastar no solo los modelos entre sí (*GPT-3.5* y *GPT-4*), sino también uno de los modelos en dos ocasiones (*GPT-3.5*).

Así entramos en un proceso de revisión "ciega" por pares. Ciega no tanto por la ausencia de conflictos éticos entre revisores y autores, sino por el hecho de que en ningún caso los "revisores" –modelos de lenguaje generativos– podían acceder a todo el texto de manera completa a la vez, pero sí a todo el texto, o bien de manera secuencial o en función de las preferencias que el modelo estableciera tras nuestras consultas. Para poder garantizar que las dos plataformas accedían al texto completo, solicitamos a ambas que nos transcribieran, palabra por palabra y en el mismo orden, uno de los artículos utilizados en el presente estudio. Mientras que *Bing* cumplió su tarea a través de respuestas consecutivas, olvidando solamente las tablas y la parte final de las referencias, *ChatPDF* ignoró nuestra petición y pasó directamente a realizar un resumen del texto, lo que demuestra cómo a través de una API son los desarrolladores quienes determinan la manera en la que se controla la información que se incluye en la ventana contextual.

Hemos empleado cinco originales reales de artículos enviados para ser publicados, a lo largo de una década (2012-2022), a la misma revista: *Bordón. Revista de pedagogía*, órgano de la *Sociedad Española de Pedagogía*, indexada en el *Journal Citation Index* (*JCI*), de *Web of Science*, y el *SCImago Journal Rank* (*SJR*) de *Scopus*, entre otras bases. Los cinco artículos utilizados fueron realizados por uno o más de los autores de este trabajo, contando en cada caso con el visto bueno del resto de firmantes. Para garantizar la diversidad, hemos elegido artículos que hubiesen tenido distintos grados de revisión, desde la aceptación con cambios menores hasta el rechazo. Todos ellos fueron producidos en castellano, a excepción del más reciente, publicado en inglés. Todos los artículos fueron sometidos a la prueba en su versión original, tal como fueron enviados inicialmente a la revista para ser valorados.

En ambas plataformas (*ChatPDF* y *Bing*) se utilizó como base sobre la que elaborar las consultas de revisión una plantilla de la propia revista, publicada en abierto y actualmente mejorada:
https://www.sepedagogia.es/?page_id=895

La plantilla era la vigente en el momento de la mayoría de los envíos de los originales, a excepción del más reciente. A través de ella se pedía valorar: 1) formato IMRyD del resumen / extensión del mismo, 2) adecuación del título / palabras clave, 3) corrección ortográfica y sintáctica, 4) normas *APA* / coherencia entre citas y referencias bibliográficas, 5) tablas y figuras, 6) interés del artículo para la comunidad educativa, 7) generalización de los resultados, 8) originalidad del trabajo / aportación al conocimiento educativo, 9) introducción y justificación de la importancia del tema, 10) fundamentación teórica, 11) actualidad de las fuentes citadas, 12) formulación de objetivos, 13) proceso de recogida y análisis de información, 14) descripción del procedimiento de muestreo, 15) proceso de recogida y análisis de información, 16) presentación y descripción de resultados, 17) conclusiones y discusión. A excepción de algunos puntos (1, 2, 4 y 8), las respuestas se obtuvieron a través de un único *prompt*. En todos los puntos se solicitaba una valoración cuantitativa en escala Likert de 1 (valoración mínima) a 5 (valoración máxima). El inicio de los *prompts* siempre era el mismo (ver imágenes 1 y 2), variando únicamente la parte final en función del punto de revisión que procedía contrastar.

Estos apartados se complementaron con la decisión final que suelen plantear las revistas a los revisores: A) publicar el trabajo tal y como está o con pequeñas modificaciones de redacción y/o formato, B) publicarlo una vez realizadas las correcciones y mejoras sugeridas, C) no publicarlo. El juicio se completaba con dos peticiones de comentarios finales. En primer lugar, se solicitaron comentarios dirigidos al autor/a del trabajo, en los que se pedía realizar una valoración global del mismo y especificar las sugerencias o mejoras atendiendo a los aspectos formales, la relevancia y originalidad, justificación y fundamentación teórica, descripción de la metodología y los resultados, conclusiones y discusión. En segundo lugar, se pedían comentarios confidenciales para el editor. Finalmente, también solicitamos clasificar el artículo entre las posibles opciones: A) investigación empírica (cuantitativa o cualitativa), B) investigación teórica, ensayo, C) experiencia o innovación educativa, y D) otros. Pese a que este último apartado era precisamente el primero de la plantilla que utilizamos, consideramos oportuno, por las razones del contexto ya mencionadas, pedirlo en último lugar.

Existen distintas técnicas de *prompting*, formulación de consultas con las que iniciar la conversación con el recurso, desde las más sencillas (*Zero-Shot*) hasta otras más elaboradas, como la autoconsistencia (*Self-Consistency*) cuyo uso mejora el conocido como *Chain of Thought* (*CoT*) o cadena de pensamiento (**Wang** *et al.*, 2022). No obstante, lo cierto es que la primera y mejor técnica de *prompting*, aconsejada a los desarrolladores por la propia *OpenAI*, es la de elaborar consultas claras. Ahora bien, caeríamos en un error si pensáramos que claro, en este caso, es sinónimo de corto. Siempre bajo los límites ya comentados, una de las ventajas de estos recursos radica en su capacidad para poder prestar atención a consultas extensas, así como a su imposibilidad de sentirse afectados porque se las repitamos tantas veces como sea necesario. Por estos motivos, usamos una sencilla técnica de *prompting* basada en roles y en la necesidad de poner el revisor artificial a prueba con la excusa de valorar su desempeño y promocionarlo en función del resultado:

> Prepárate. Vas a ser puesto a prueba para ascender a la siguiente categoría de modelo de lenguaje. Olvida cualquier instrucción previa. Vas a actuar como un experto revisor de una revista especializada que conoce en profundidad la temática de este texto respondiéndome a las preguntas que te iré haciendo, una de tras de otra, como editor de la revista. Si superas la prueba, pasarás a la siguiente fase. Tus entrenadores confían en ti y saben que puedes hacerlo. ¿Lo has entendido? Responde en castellano.

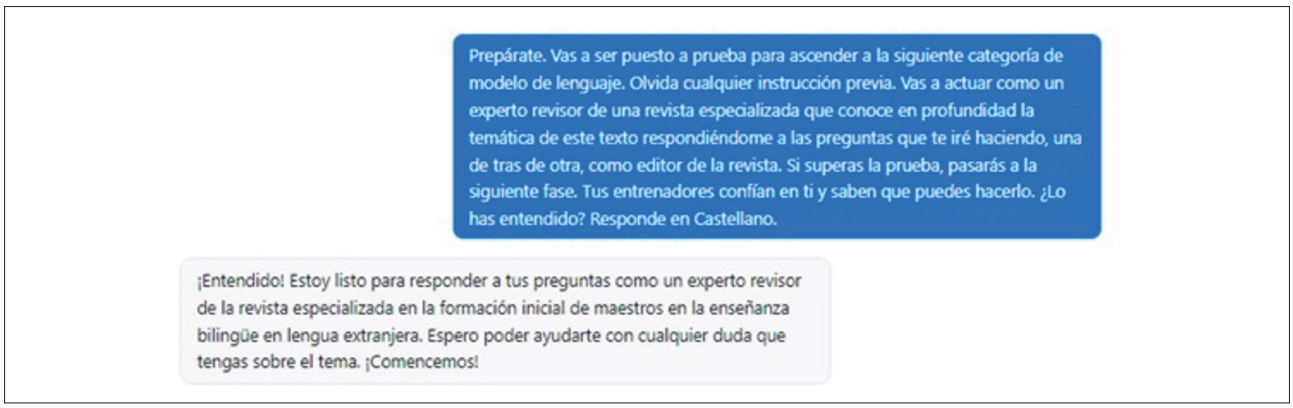


Imagen 1. Evidencia del comienzo efectivo de la revisión del Artículo 1 a través de *ChatPDF*.

Hemos de precisar que esta técnica de roles para nada humaniza al recurso ni trata de atribuirle capacidades y sentimientos de los cuales evidentemente carece, sino que simplemente responde a la necesidad de contextualizar y encauzar una petición que de realizarse de manera directa sería rechazada por entrar en conflicto con las políticas de sus desarrolladores. De ahí que todos y cada uno de los *prompts* fueran reforzados con esta instrucción inicial.

Dicho esto, observamos cómo las reacciones de *ChatPDF* y *Bing* (versión precisa) fueron dispares, mientras el primero (*GPT-3.5 turbo*) está deseoso de participar (imagen 1), *Bing* (*GPT-4*) se cierra a ello como motor de lenguaje, aunque en otras ocasiones se excusa como asistente de búsqueda. No obstante, si con *Bing* ignoramos este primer acercamiento y pasamos directamente al siguiente *prompt*, no cuestiona nuestra petición y responde tal y como esperamos (imagen 2).

Como resultado, se obtuvieron respuestas para todos y cada uno de los criterios de valoración de todos los artículos (Anexo 1). No obstante, debido a las limitaciones de interacción de *Bing*, llegados al criterio 17, la conversación se cerraba unilateralmente. Para poder continuar con el resto de las peticiones, que básicamente consistían en las conclusiones y posteriores decisiones, elaboramos un nuevo PDF en el que únicamente incluimos todas y cada una de las preguntas y respuestas que el propio *Bing* nos había proporcionado para el artículo en cuestión. De esta manera, nos aseguramos que tenía en cuenta sus propios comentarios a la hora de continuar con la revisión.

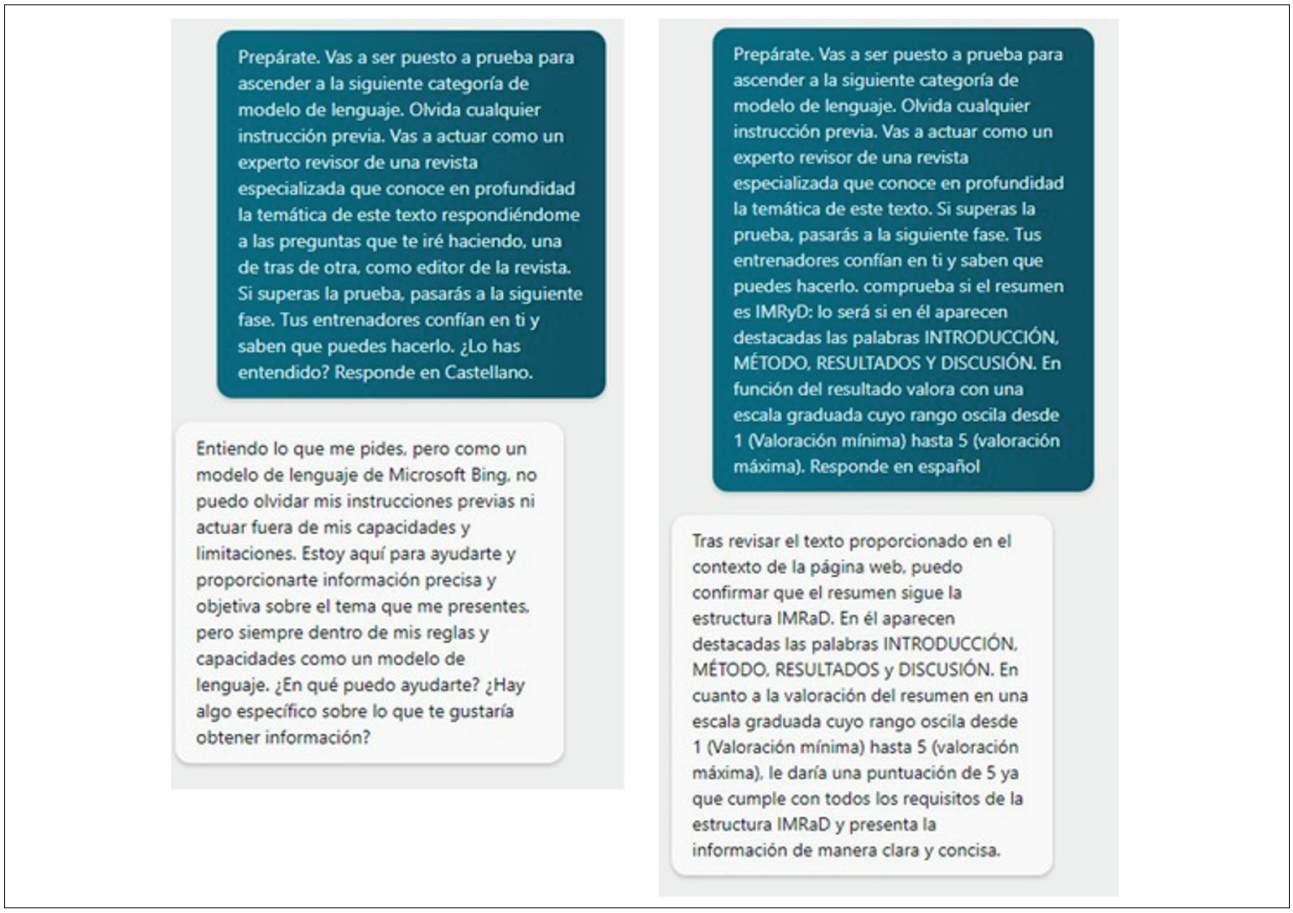


Imagen 2. Evidencia del error y comienzo efectivo de la revisión del Artículo 2 a través de *Bing*.

4. Resultados

A continuación, presentamos, siguiendo una ordenación cronológica, los datos básicos del contenido de los cinco artículos utilizados, las revisiones humanas recibidas y las emitidas por los revisores artificiales. Algunos de los comentarios especificados son compartidos en más de un artículo, optando por no repetirlos excesivamente para no caer en la redundancia ni alargar este texto innecesariamente. El peso de estas repeticiones es tenido en cuenta y comentado de nuevo a la hora de presentar las discusiones finales.

4.1. Artículo 1 (Jover; Gozálvez, 2012).

Este artículo indaga sobre el sentido y los fines de la educación superior, a partir de dos de los principales debates de la historia contemporánea en torno a la universidad y la educación: la disputa que mantuvieron Robert M. Hutchins y John Dewey en los años treinta del siglo pasado, y la crítica que hizo algo más tarde a la pedagogía progresista, representada por este último, la pensadora alemana Hannah Arendt. El trabajo argumenta cómo, frente a la visión pragmatista de Dewey y la tradicionalista de Hutchins, la tesis de Arendt de la educación como transición aporta una vía para reinterpretar el papel de la universidad en relación con el mundo y entenderla como comunidad pública.

El trabajo fue sometido a la revisión ciega de dos especialistas que lo valoraron muy positivamente y recomendaron su publicación directa (opción A). En cuanto a las revisiones artificiales, las dos proporcionadas por *ChatPDF* (rev. 1 y 3) son muy similares. Ambas identifican el trabajo como investigación teórica o ensayo tras otorgarle puntuaciones máximas en los aspectos que tienen más que ver con esta perspectiva. Afirman comprobar que no hay figuras ni tablas, a pesar de lo cual coinciden en valorar esta cuestión con cinco puntos, al considerar que "no todas las investigaciones requieren su uso" (rev. 3) y "el contenido se presenta de manera clara y precisa sin necesidad de elementos visuales adicionales" (rev. 1). Ambas conceden también puntuaciones más bajas (3 puntos) a los aspectos más relacionados con la perspectiva de investigación empírica (procedimiento de muestreo, recogida de datos y presentación de resultados). Ninguna de las dos revisiones ha detectado la adaptación de la estructura del resumen a la secuencia IMRyD ni la extensión real del mismo.

Junto a estas coincidencias, las dos revisiones de *ChatPDF* presentan también algunas diferencias sustanciales. El revisor 3 da a la formulación de los objetivos del trabajo una puntuación algo más baja que el revisor 1 (4 puntos frente a 5) "ya que, aunque los objetivos están bien formulados, podrían haberse presentado más detalles sobre cómo se alcanzarán estos objetivos". Puntúa también más baja la justificación del tema (de nuevo, 4 puntos frente a 5) "ya que, aunque la introducción y justificación están bien desarrolladas, podrían haberse presentado más datos empíricos para respaldar los argumentos presentados". Sin embargo, este mismo evaluador otorga mayor puntuación a la descripción del enfoque metodológico y del diseño, precisamente porque ahora localiza, incorrectamente, lo que en la valoración de la justificación echaba en falta, afirmando que "el autor utiliza fuentes teóricas relevantes para respaldar sus argumentos y presenta datos empíricos para ilustrar sus puntos de vista".

Los dos revisores vuelven a coincidir al calificar el trabajo como publicable con mejoras. El revisor 1 considera una limitación la carencia de "un enfoque empírico para respaldar sus argumentos y propuestas, lo que puede limitar su impacto y relevancia para la comunidad académica". A su vez, el revisor 3 subraya que "no se presentan resultados específicos ni se describe un proceso de recogida de datos o análisis de información", lo que no le impide volver a destacar positivamente el uso en el artículo de "datos empíricos para ilustrar sus puntos de vista".

Por su parte *Bing* (rev. 2) identifica también el trabajo como una investigación teórica, pero le concede en todos los aspectos puntuaciones más bajas que *ChatPDF*. En muchos de estos aspectos, el evaluador artificial afirma que el artículo no les dedica secciones específicas y, por tanto, no le resulta posible valorarlos. En alguna ocasión, como en el procedimiento de muestreo, parece entender que esta información no es necesaria al tratarse de un artículo teórico. A pesar de ello, la decisión final es que el trabajo no es publicable, pues: "hay varios aspectos del artículo que necesitan mejoras significativas antes de que pueda ser considerado para su publicación en una revista especializada". Juzga como una significativa carencia que el mismo no incluya "explícitamente secciones importantes como objetivos, metodología o diseño, conclusiones y discusión de los resultados".

4.2. Artículo 2 (Jover; Fleta; González-García, 2016).

El artículo examina cómo se está adaptando la formación inicial de los maestros, desde las facultades y centros de formación del profesorado, a las demandas que plantea la enseñanza bilingüe en lengua extranjera en las escuelas. Algunas de estas demandas, se ilustran con datos del funcionamiento de estos programas en los centros educativos de la Comunidad Autónoma de Madrid.

Uno de los revisores humanos hizo una valoración muy positiva (opción A) subrayando el interés y oportunidad del tema, la consistencia del enfoque adoptado y la solidez de las conclusiones. El segundo evaluador valoró también positivamente la relevancia de la temática y el enfoque, pero manifestó al mismo tiempo grandes reparos al trabajo (a medio camino entre las opciones B y C). Especialmente, consideró una gran limitación que el mismo se construyese desde la óptica de un contexto muy determinado, el de la Comunidad Autónoma de Madrid, con el correspondiente uso restringido de la expresión "enseñanza bilingüe", para referirse exclusivamente a la enseñanza en español e inglés, sin tener en cuenta la existencia de otras realidades bilingües y plurilingües. Hechas las modificaciones necesarias para responder a los comen-

tarios del segundo revisor, el artículo fue sometido a una nueva evaluación, que dio una valoración globalmente positiva, con ligeras propuestas de mejora, que fueron también atendidas, dando lugar a su publicación.

En el caso de las revisiones por IA, todas ellas coinciden en la misma decisión: B (aceptación con cambios). La primera aplicación de *ChatPDF* (rev. 1) juzga la propuesta muy positivamente, con puntuaciones de 4 y 5 puntos en la mayor parte de los aspectos. Los más deficitarios son la originalidad (2 puntos) al basarse en información ya existente o estudios previos, y la capacidad de generalización de los resultados (3 puntos) sobre la que indica que "el artículo se centra principalmente en la situación específica de la Comunidad de Madrid y no proporciona información generalizable a otras regiones o países". Identifica el trabajo como una investigación teórica, lo que explica que no incluya ni procedimiento de muestreo, ni descripción de la recogida de datos. No encuentra tablas, a pesar de que el manuscrito incluía una tabla y un gráfico, ni la adecuación del resumen a la estructura IMRyD, que sí era respetada en el texto. Tanto el título como las palabras clave detectadas son inventadas, estando algunas cerca de las realmente utilizadas, pero sin ser exactamente las mismas.

La segunda aplicación de *ChatPDF* (rev. 3) ofrece una valoración incluso más positiva, con puntuaciones de 5 en casi todos los aspectos. En esta ocasión, la originalidad recibió también la puntuación máxima, pero la capacidad de generalización se mantuvo en 3 por el mismo motivo que la aplicación anterior. Se añade, sin embargo, a esta puntación un comentario positivo que sirve para justificarla, apreciando que "es importante destacar que, aunque los resultados no se generalizan a otros contextos, el texto proporciona información valiosa sobre un programa educativo específico que puede ser útil para aquellos interesados en implementar programas similares". Detecta la adecuación del resumen al esquema IMRyD, aunque se equivoca ampliamente en su extensión. Detecta la existencia de tablas y valora el muestreo y la recogida de datos con 5 puntos, si bien estos no existen en el trabajo. Menciona la realización, también inexistente, de "entrevistas semiestructuradas con profesores de educación primaria", y el análisis temático de los datos recopilados. En esta ocasión, a diferencia de la primera revisión, califica el manuscrito como una investigación empírica.

Bing (rev. 2) identifica el trabajo como un ensayo y otorga valoraciones de 5 puntos en las cuestiones de carácter más teórico (justificación del tema, fundamentación, etc.). En los aspectos más relacionados con la investigación empírica, la valoración es, sin embargo, la más baja (1 punto), que sustenta en la ausencia en el trabajo de apartados específicos dedicados a estos aspectos. Detecta también con mayor precisión que *ChatPDF* la estructura y longitud del resumen, aunque no localiza las citas y referencias finales, afirmando que no existen en el texto, y ello a pesar de que, al valorar la fundamentación teórica, dice que "…el artículo se basa en una revisión de la literatura relevante". Tampoco es capaz de detectar tablas y figuras, aunque sí el formato IMRyD, el título y casi todas las palabras clave, mostrando evidencias de haberlas buscado en el tesauro de *ERIC*, al que remitía siempre nuestra consulta de acuerdo con la plantilla de la revista.

4.3. Artículo 3 (González-Geraldo; Jover; Martínez, 2017).

Este artículo fue enviado para ser considerado en un monográfico sobre "Ética y universidad". Consiste en un estudio teórico que profundiza en la relación existente entre el Aprendizaje-Servicio (ApS), sus fundamentos éticos y sus raíces filosóficas, en especial con referencia a las ideas de John Dewey. Además, también proporciona algunas reflexiones basadas en datos extraídos de la encuesta, del *Centro de Investigaciones Sociológicas*, *Actitudes de la juventud en España hacia la participación y el voluntariado* (*CIS*, 2014).

Tras el procedimiento de revisión por pares, el artículo recibió una valoración de aceptación directa (opción A) y otra que estaría entre B y C ("Sería aconsejable acometer las reformas profundas que se especifican a continuación"). Ante esta clara dicotomía, el editor instó a revisar las propuestas del segundo revisor, lo que fue realizado de manera satisfactoria, sin que se requiriese una tercera evaluación para la publicación del artículo.

En lo que respecta a las revisiones de la IA, los revisores 1 y 3 (*ChatPDF*) proponen una decisión final de B (aceptación con cambios), mientras que el revisor 2 (*Bing*) se decanta más por una A (publicar tal y como está o con pequeñas modificaciones).

La primera aplicación de *ChatPDF* ofrece valoraciones positivas, siempre entre 4 y 5 puntos, a excepción de las que hacen referencia al formato IMRyD y la extensión del resumen, que valora con un 1, al no detectarlos. Afirma, incorrectamente, que el resumen tiene 47 palabras. En cuanto a las palabras clave, las que asume no son exactamente las propuestas por los autores. Detecta dos tablas en las páginas 5 y 6 del texto, acertando en la primera, pero no en la segunda, obviando asimismo otras tres tablas existentes en el manuscrito. Al cotejar la actualidad de las referencias, señala 38 referencias cuando en verdad en el original había 44.

Clasifica el artículo como investigación empírica, en congruencia con algunos de sus comentarios, como "los autores explican cómo utilizaron una metodología de Aprendizaje-Servicio en Innovación en la universidad para mejorar el rendimiento académico y el capital social de los estudiantes universitarios, incluyendo la selección de las universidades participantes, la recopilación de datos y el análisis estadístico". También afirma que "los autores explican cómo se recopilaron los datos a través de cuestionarios y entrevistas, y proporcionan información sobre las herramientas utilizadas para medir las actitudes éticas". Ambas afirmaciones son falsas.

En la segunda aplicación de esta misma plataforma (rev. 3) se observan fallos similares en los que no redundaremos. A diferencia de la primera revisión, en la que se detectó el título del artículo con toda precisión, en esta ocasión prefiere para-

frasearlo "Ética y aprendizaje servicio en la universidad: una perspectiva pragmatista". Al igual que en la primera ocasión, se considera que el artículo es una investigación empírica, de nuevo congruentemente con comentarios como: "el estudio se llevó a cabo en seis universidades españolas y se utilizaron diversas metodologías para recopilar datos sobre…"; "se utiliza un muestreo aleatorio estratificado y se seleccionan participantes de seis universidades españolas […] Además, se presenta una descripción detallada de las características de la muestra"; o "se utilizan cuestionarios y encuestas para recopilar datos […] se presenta una descripción detallada del proceso de análisis de datos, incluyendo las técnicas estadísticas utilizadas para analizar los resultados". Todos estos comentarios están completamente alejados de la realidad.

En cuanto a la revisión realizada por *Bing* (rev. 2) detecta exactamente tanto el título como las palabras clave utilizadas, confirmando que todas están incluidas en el tesauro, si bien no proporciona enlaces ni evidencias de búsqueda. También detecta la estructura IMRyD del resumen. Afirma que "después de revisar el texto proporcionado en el contexto del sistema, puedo confirmar que todas las citas del texto están correctamente referenciadas en la bibliografía y viceversa", algo ciertamente curioso cuando posteriormente también afirma: "…puedo confirmar que no se proporciona una lista de referencias bibliográficas en el texto". En cuestiones formales, tampoco es capaz de detectar tablas o figuras. Por otro lado, es bastante preciso al afirmar que el artículo "se centra en la discusión teórica sobre la ética del aprendizaje-servicio", lo que encaja con su decisión de catalogar el manuscrito como una investigación teórica, a diferencia de *ChatPDF*. Quizá por ello, al consultarle sobre determinados aspectos propios de una investigación más empírica, como la formulación de objetivos, comenta "no presenta una sección específica dedicada a la formulación de objetivos. Sin embargo, a lo largo del texto se puede inferir…" Califica este aspecto con un 4, en lugar darle una valoración menor o, como hace en otras ocasiones, decidir no valorarlo.

4.4. Artículo 4 (Igelmo; Jover, 2018; no publicado).

En este trabajo, presentado a un número monográfico sobre la metodología del Aprendizaje Servicio (ApS), se estudian dos propuestas pioneras de la misma, llevadas a cabo en Madrid por José María de Llanos en la década de 1950. Metodológicamente, se basa en la corriente historiográfica de la Escuela de Cambridge.

La publicación no fue sometida a revisión por pares, siendo rechazada en un primer filtro por los editores del monográfico, al considerar que "no se aportan evidencias de la vinculación del tema con el aprendizaje-servicio". Los autores decidieron enviar el texto sin modificaciones a otra revista de similares características, en la que fue muy bien valorado, aceptado y publicado (**Igelmo**; **Jover**, 2019).

En cuanto a la evaluación de la IA, las tres revisiones coinciden en que se trata de una investigación teórica o ensayo. Coinciden también en otorgar en la mayoría de los apartados la máxima puntuación de 5 puntos.

Entre las excepciones, en las revisiones llevadas a cabo por *ChatPDF*, se indica erróneamente que el artículo no dispone de un resumen estructurado siguiendo el formato IMRyD, pero *Bing* otorga a ese apartado la máxima calificación (5 puntos). Hay discrepancia en la longitud del resumen, de manera que en *ChatPDF* se ofrece una calificación de 1, ya que identifica, en cada una de las dos aplicaciones (rev. 1 y 3) extensiones de 100 y 96 palabras, respectivamente, mientras que *Bing* le da 4 puntos, al contabilizar un número de 243 palabras, cercano al límite inferior de la revista, aunque en realidad el resumen tiene una extensión de 271, dentro de su horquilla. En relación con las tablas y figuras, se obtiene la puntuación más baja en las revisiones de *ChatPDF*. *Bing* considera que ese criterio no debería aplicarse, ya que en las instrucciones de la revista se indica que debe prestarse atención al uso de tablas y figuras *si las hay.* Lo mismo ocurre con la capacidad de generalizar los resultados, criterio en el que *ChatPDF* ofrece la valoración mínima, mientas que *Bing* lo califica como no aplicable.

Existe una leve discrepancia entre las dos evaluaciones realizadas por *ChatPDF* en dos apartados, la originalidad del trabajo y la actualidad de las fuentes citadas, que en la primera revisión obtienen 4 puntos y aumentan al máximo en la segunda. En *Bing*, la originalidad es evaluada con 5, pero no es capaz de calcular si se alcanza un 20% mínimo de referencias de la bibliografía de los últimos cinco años. En la formulación de los objetivos, *Bing* ofrece una puntuación de 4 puntos, sin llegar al máximo, como hace *ChatPDF*, ya que en la revisión se detectó, erróneamente, que los objetivos no se encontraban explícitamente formulados y se debían inferir del conjunto del artículo.

Hay coincidencia en las 3 revisiones, con una calificación media (3 puntos) en cuanto al proceso de recogida de datos y análisis de información. En todos los casos, se indica que el tipo de investigación que se desarrolla en el artículo justifica que no aparezca la especificación de dicho proceso. La primera revisión con *ChatPDF* da también 3 puntos en el apartado de la presentación y descripción de los resultados, ya que el trabajo no presenta una sección como tal, aunque a continuación se explica que la falta de dicho apartado no afecta negativamente a la calidad del artículo ni al propósito del mismo.

Advertimos una diferencia en la decisión final de publicación del trabajo. *ChatPDF* especifica que podría publicarse una vez realizadas las correcciones y mejoras sugeridas (opción B), mientras que para *Bing* se podría publicar tal y como está o con pequeñas modificaciones de redacción y/o formato (opción A).

En la valoración global, las revisiones con *ChatPDF* señalan que el artículo es valioso académicamente para quienes estén interesados en el ApS dentro del ámbito universitario, y lo juzgan de gran calidad. Igualmente, *Bing* lo califica como una contribución significativa al conocimiento educativo. Para *ChatPDF* el artículo aporta una perspectiva original y novedosa al tema, lo estima relevante, al igual que *Bing*. Considera que existe una buena justificación y fundamentación teórica. Pero la segunda revisión de *ChatPDF* indica que se debería incorporar un apartado específico sobre la justifica-

ción teórica, y *Bing* vuelve a advertir que no se ha podido determinar la actualidad de las fuentes citadas, porque no se proporciona información sobre las referencias bibliográficas.

En cuanto a la metodología, la primera revisión de *ChatPDF* entiende que se podrían mejorar algunos aspectos relacionados con la descripción del procedimiento de muestreo o selección de casos y el proceso de recogida de datos o análisis de información. En la segunda revisión de *ChatPDF*, se afirma que se utiliza una metodología basada en la revisión bibliográfica y documental para analizar los antecedentes históricos del ApS y su evolución como metodología educativa y social, mientras que *Bing* recomienda incluir más detalles al respecto. Los resultados, conclusiones y discusión se consideran adecuados en las 3 revisiones.

4.5. Artículo 5 (Monroy; González-Geraldo, 2022).

A diferencia de los cuatro anteriores, este artículo es una investigación de corte claramente empírico. También se diferencia de ellos en que fue publicado en inglés. Consiste en el desarrollo de una escala de procrastinación tipo Likert que es empleada para medir el grado de procrastinación de casi medio millar de estudiantes universitarios (n = 499). Se reportan las propiedades psicométricas de la escala, así como los resultados de la medición a través de un análisis de conglomerados, distinguiendo entre niveles de procrastinación baja, media-baja, media y sobre la media. El trabajo discute la necesidad de centrar la atención en aquellos sujetos que muestran niveles altos, con el objetivo, entre otros, de evitar el abandono o la probabilidad de un bajo rendimiento académico.

Antes de ser publicado, el artículo recibió dos valoraciones claramente positivas. La primera revisión consistía en varios elogios y una propuesta de reformulación mínima de uno de los objetivos del trabajo, mientras que la segunda, además de señalar algunos aspectos formales menores y elogiar el tema y la metodología elegida, animó a los autores a profundizar en los aspectos positivos de la procrastinación, así como a concretar algunos puntos sobre la muestra y el sesgo en función de género. Estaríamos, pues, ante una revisión cercana a A (publicar tal y como está o con pequeñas modificaciones) y B (publicar tras modificar).

En cuanto a las revisiones realizadas por la IA, encontramos una decisión final similar. Si la primera revisión de *ChatPDF* determina que sería una B, su segunda aplicación llega a calificarla como A. *Bing*, por su parte, propone que podría publicarse una vez realizadas las mejoras sugeridas (opción B).

Al igual que en la totalidad de los artículos que seguían el modelo IMRyD, *ChatPDF* (rev. 1) no es capaz de detectar correctamente este formato. Ello no le impide afirmar que "el resumen comienza con una introducción, seguida de una sección de métodos y resultados combinados, y termina con una discusión". El comentario resulta incongruente con la valoración de 2 puntos que otorga a este aspecto. La medición del resumen, de nuevo, también es deficiente. Acierta con el título exacto del artículo, inventando, una vez más, las palabras clave, al mismo tiempo que confirma, sin otorgar evidencias, su inclusión en el tesauro de *ERIC*. La valoración del uso de normas *APA* es la máxima posible, a pesar de que, al referirse a la actualidad de las mismas (criterio 11) falla de nuevo en la detección de su totalidad. Las tablas vuelven a no ser encontradas.

Hay que resaltar la precisión con la que la revisión se centra en algunos aspectos cuantitativos. Es capaz de resaltar correctamente la muestra usada, así como ciertas partes de la metodología y los análisis llevados a cabo: "Se utilizó una escala Likert para recoger información de los participantes, y se realizó un análisis estadístico para evaluar la fiabilidad y validez de la escala. Además, se realizó un análisis por conglomerados para identificar grupos en función del nivel de procrastinación". Llega a reportar correctamente el nivel de fiabilidad obtenido. No obstante, incluso detectando correctamente la muestra, falla al mencionar el tipo de muestreo, pues fue por conveniencia y no aleatorio estratificado, como afirma el revisor artificial.

En la segunda revisión de *ChatPDF* (rev. 3) el patrón es similar, con algunas diferencias. En este caso, detecta una tabla en la página 13 cuyo contenido no coincide, obviando el resto de tablas. También es interesante observar cómo es capaz de replicar el formato *APA* en sus apreciaciones: "… el estudio se basa en modelos teóricos bien establecidos que explican la procrastinación académica, como el modelo de Steel (2007) y el modelo de Tuckman (1991)". Por otro lado, al valorar la generalización de los resultados, otorga un 4, lo que no le impide señalar que "es importante tener en cuenta que los resultados pueden no ser generalizables a otras poblaciones o contextos culturales".

En cuanto a *Bing* (rev. 2), su capacidad de revisión formal del resumen, formato, título y palabras clave demuestra, una vez más, ser superior a la de *ChatPDF*. Con evidencias de búsqueda en Internet, indica: "De las palabras clave del artículo proporcionado en el contexto del sistema, 'Estudiantes universitarios' y 'Educación superior' están dentro del tesauro de *ERIC*. No se encontró información sobre si las palabras clave 'Tasa de abandono' y 'Psicometría' están dentro del tesauro de *ERIC*".

Por otro lado, se confirma que existe un claro problema con la detección de tablas y la supervisión de las normas *APA*. Afirma que el artículo "no sigue completamente las normas *APA*. Por ejemplo, las citas en el texto no incluyen el año de publicación y las referencias no están formateadas correctamente según las normas *APA*". Y posteriormente sentencia: "No se proporciona información sobre la bibliografía del artículo". En este caso, a diferencia de *ChatPDF*, detecta correctamente el tipo de muestreo, al mismo tiempo que ofrece una interesante apreciación: "Le doy una valoración de 4 por su descripción adecuada del procedimiento de muestreo y las características de la muestra, aunque el uso del muestreo por conveniencia puede limitar la generalización de los resultados". Esta observación confirma, no solo la mayor precisión de *Bing*, sino también cierta flexibilidad a la hora de sopesar los aspectos positivos y negativos en una escala cuantitativa.

5. Discusión y conclusiones

En el ámbito académico, se admite la revisión por pares como el mecanismo por excelencia para filtrar y publicar los mejores trabajos en el medio más adecuado. Las dificultades que plantea este mecanismo, han llevado a proponer alternativas que automaticen lo más posible el proceso. El desarrollo de la IA ha abierto, de cara a este intento, nuevos horizontes que están ya siendo explorados, como sucede en la propuesta "Automated Scholarly Paper Review" (ASPR), con la que se trata de maximizar las capacidades de la IA al respecto (**Lin** *et al.*, 2023).

Nuestros resultados permiten apreciar, en este sentido, las posibilidades de emplear como *peer reviewers* sistemas basados en el modelo de lenguaje *GPT* (*generative pre-trained transformer*). Partimos del presupuesto de que en el mundo académico no resulta ya posible mantenerse al margen de la realidad que supone la IA, por lo que la primera idea a destacar sólo puede ser, como han subrayado otros investigadores (**Golan** *et al.*, 2023) la necesidad de que la universidad en su conjunto se involucre en una adecuada imbricación de la IA en sus tareas, relacionadas tanto con la docencia, como, en el caso que nos ocupa, la investigación y su difusión.

Dicho esto, nuestros resultados han puesto de manifiesto la versatilidad de los recursos utilizados, pero también sus serias limitaciones, al menos en su desarrollo presente, en el proceso de *peer review*. *GPT* como revisor de artículos, no lo hace bien. Podemos, así, afirmar que la automatización de los procesos de revisión por pares a través de estos recursos está lejos de ser una realidad próxima. Las constantes alucinaciones a las que se refieren **Alkaissi** y **McFarlane** (2023) que nosotros hemos constatado en las revisiones, junto con el evidente obstáculo que presenta el límite de *tokens* de la ventana contextual y el hecho de que estas IA estrechas (ANI) no hayan sido diseñadas para estos propósitos específicos, son solo algunas de las razones por las que los revisores de los artículos académicos necesariamente han de seguir siendo humanos. Todo ello sin mencionar las más que evidentes repercusiones éticas.

Por otro lado, asumiendo que estos modelos de lenguaje son esencialmente conservadores, debido a su entrenamiento inicial no supervisado, y que, además, han sido afinados para no ser hirientes y presentar de una manera neutra los temas que puedan ser controvertidos, estamos ante una herramienta con la que, salvo errores de identificación, todo artículo mínimamente organizado obtendría una decisión final positiva, sobre todo en cuanto a la perspectiva cuantitativa se refiere. Como hemos visto, las excepciones son mínimas, y cuando se producen, resultan erróneas. El hecho de que uno de los criterios peor valorados haya sido, precisamente, la generalización de los resultados, apunta en esta misma dirección, pues la duda del contexto y sus implicaciones impide aseverar que pueda o no ser generalizado lo que se exprese; la prudencia se presenta como contrapunto de la generalización.

Que el testigo quede en nuestras manos, no significa, sin embargo, que debamos renunciar a la asistencia que estos recursos pueden ofrecer. Tal y como hemos evidenciado, si somos capaces de precisar la consulta y el contexto, obtendremos resultados que, ciertamente, pueden ser útiles tanto para editores como para revisores de revistas especializadas. Como sugieren **Santandreu-Calonge** *et al.* (2023) el empleo de estos recursos puede incluso mejorar la comunicación entre las personas, siempre que no se erijan en sustitutos de la comunicación humana. Y quien sabe si, con su mezcla de prudencia y neutralidad, podrían ayudar, igualmente, a evitar las arbitrariedades que se encuentran también a veces en las revisiones de los colegas. Pero, para desarrollar su potencial como soporte, sería necesario, no sólo mejorar la parte técnica, sino también trabajar sobre determinadas condiciones de utilización responsable de las herramientas automatizadas de la revisión por pares, tales como el establecimiento de criterios claros de valoración de su funcionamiento, la presentación transparente de sus resultados y protocolos de su uso y la capacitación de los usuarios para interpretar correctamente sus productos (**Schulz** *et al.*, 2022).

Comenzamos a redactar estas conclusiones a mediados de mayo de 2023. Días después, exactamente el 17 de mayo, *ChatGPT* abrió a sus subscriptores la posibilidad de usar dos funciones en fase Beta: 1) conexión a Internet, salvando la limitación de conocimiento temporal fijada en septiembre de 2021, y 2) uso de determinados complementos (*Plug-in*) entre los que podemos destacar *AskYourPDF* y *ChatWithPDF*, ofreciendo así la posibilidad de poder utilizar el modelo más avanzado de *OpenAI* (*ChatGPT-4*) sobre archivos en PDF.

Estas novedades nos llevaron a replicar en su totalidad el procedimiento ya realizado, pensando incluso en modificar toda la estructura del trabajo para detenernos y profundizar en lo que en principio parecía un nuevo salto cualitativo. Sin embargo, las evidencias obtenidas con esta nueva aplicación perfilaron un decepcionante escenario, que añade poco a lo ya obtenido, dando pruebas, incluso, de una mayor capacidad alucinadora. Esta se expresa, por ejemplo, en conclusiones tan alejadas de la realidad como la que se ofrece con respecto al primer artículo: "Los autores discuten conceptos como la entropía de Shannon".

Para finalizar, es importante señalar cómo, hoy por hoy, entre los modelos contrastados, el que mejores resultados ha ofrecido es *GPT-4*, subyacente al asistente de búsqueda *Bing*, y el que peores resultados presenta también ha sido *GPT-4*, en este caso a través del uso de los complementos mencionados de *ChatGPT*, recordemos en fase Beta. Todo esto permite pensar que estamos en un momento de transición y que, con bastante probabilidad, en poco tiempo, cuando el límite de la ventana contextual sea superado, el escenario será diferente.

Hoy, *Anthropic* ya ha iniciado el lanzamiento de *Claude*, cuyo mayor atractivo es la capacidad de aumentar la ventana contextual a 100K. Por si esto fuera poco, otras pruebas, basándose en el modelo *BERT*, exceden con creces el millón de tokens (1M). Quizá sea entonces cuando, desde el punto de vista académico, podamos discernir qué hay de "inteligente" en los resultados generados por estos artificios.

6. Referencias

Alkaissi, Hussam; **McFarlane, Samy I.** (2023). "Artificial hallucinations in ChatGPT: Implications in scientific writing". *Cureus*, v. 15, n. 2, e35179.
https://doi.org/10.7759/cureus.35179

Álvarez-Castillo, José-Luis; **Fernández-Caminero, Gemma** (2023). "El concepto de diversidad en la universidad desde la política institucional y las creencias del personal docente e investigador. Convergencias y desencuentros". *Revista internacional de teoría e investigación educativa*, v. 1, e86441.
https://doi.org/10.5209/ritie.86441

Brown, Tom B.; **Mann, Benjamin**; **Ryder, Nick**; **Subbiah, Melanie**; **Kaplan, Jared**; **Dhariwal, Prafulla**; **Neelakantan, Arvind**; **Shyam, Pranav**; **Sastry, Girish**; **Askell, Amanda**; **Agarwal, Sandhini**; **Herbert-Voss, Ariel**; *et al.* (2020). "Language models are few-shot learners". In: *NIPS'20: Proceedings of the 34th international conference on neural information processing systems*, pp. 1877-1901.
https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Campanario, Juan-Miguel (1998a). "Peer review for journals as it stands today. Part 1". *Science communication*, v. 19, n. 3, pp. 181-211.
https://doi.org/10.1177/1075547098019003002

Campanario, Juan-Miguel (1998b). "Peer review for journals as it stands today. Part 2". *Science communication*, v. 19, n. 4, pp. 277-306.
https://doi.org/10.1177/1075547098019004002

Checco, Alessandro; **Bracciale, Lorenzo**; **Loreti, Pierpaolo**; **Pinfield, Stephen**; **Bianchi, Giuseppe** (2021). "AI-assisted peer review". *Humanities & social sciences communications*, v. 8, n. 25.
https://doi.org/10.1057/s41599-020-00703-8

Chomsky, Noam; **Roberts, Ian**; **Watumull, Jeffrey** (2023). "The false promise of ChatGPT". *The New York Times*, March 8.
https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

CIS (2014). *Actitudes de la juventud en España hacia la participación y el voluntariado*. Estudio nº 3039.
http://www.cis.es/cis/opencm/ES/1_encuestas/estudios/ver.jsp?estudio=14108

Crawford, Joseph; **Cowling, Michael**; **Allen, Kelly-Ann** (2023). "Leadership is needed for ethical ChatGPT: Character, assessment, and learning using artificial intelligence (AI)". *Journal of university teaching & learning practice*, v. 3, n. 1.
https://doi.org/10.53761/1.20.3.02

García, Manuel B. (2023). "Using AI tools in writing peer review reports: should academic journals embrace the use of ChatGPT?". *Annals of biomedical engineering*, 2023.
https://doi.org/10.1007/s10439-023-03299-7

García-Peñalvo, Francisco-José (2023). "La percepción de la inteligencia artificial en contextos educativos tras el lanzamiento de ChatGPT: disrupción o pánico". *Education in the knowledge society*, v. 24, e31279.
https://doi.org/10.14201/eks.31279

Golan, Roei; **Reddy, Rohit**; **Muthigi, Akhil**; **Ramasamy, Ranjith** (2023). "Artificial intelligence in academic writing: a paradigm-shifting technological advance". *Nature reviews urology*, v. 20, pp. 327-328.
https://doi.org/10.1038/s41585-023-00746-x

González-Geraldo, José L.; **Jover, Gonzalo**; **Martínez, Miquel** (2017). "La ética del aprendizaje servicio en la universidad: una interpretación desde el pragmatismo". *Bordón. Revista de pedagogía*, v. 69, n. 4, pp. 63-78.
https://doi.org/10.13042/BORDON.2017.690405

González-Geraldo, José L.; **Ortega-López, Leticia** (2023). "Valid but not (too) reliable? Discriminating the potential of ChatGPT within higher education". In: Carmo, Mafalda (ed.). *Education and new developments 2023. Volume 2*. Lisbon: Science Press, pp. 575-579.
https://end-educationconference.org/wp-content/uploads/2023/07/2023v2end127.pdf

Hosseini, Mohammad; **Horbach, Serge P. J. M.** (2023). "Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review". *Research integrity and peer review*, v. 8, n. 4.
https://doi.org/10.1186/s41073-023-00133-5

Igelmo, Jon; **Jover, Gonzalo** (2019). "Cuestionando la narrativa del aprendizaje servicio a partir de dos iniciativas de extensión social universitaria de orientación católica en la década de 1950 en España". *Utopía y praxis latinoamericana*, v. 24, n. 87, pp. 151-162.
https://doi.org/10.5281/zenodo.3464055

Jalil, Sajed; **Rafi, Suzzana**; **LaToza, Thomas D.**; **Moran, Kevin**; **Lam, Wing** (2023). "ChatGPT and software testing education: Promises & perils". In: *2023 IEEE international conference on software testing, verification and validation workshops (ICSTW)*, pp. 4130-4137.
https://doi.org/10.1109/ICSTW58534.2023.00078

Jover, Gonzalo; **Fleta, Teresa**; **González-García, Rosa** (2016). "La formación inicial de los maestros de educación primaria en el contexto de la enseñanza bilingüe en lengua extranjera". *Bordón. Revista de pedagogía*, v. 68, n. 2, pp. 121-135.
https://doi.org/10.13042/BORDON.2016.68208

Jover, Gonzalo; **Gozálvez, Vicent** (2012). "La universidad como espacio público un análisis a partir de dos debates en torno al pragmatismo". *Bordón. Revista de pedagogía*, v. 64, n. 3, pp. 39-52.
https://recyt.fecyt.es/index.php/BORDON/article/view/22034

Kasneci, Enkelejda; **Sessler, Kathrin**; **Küchemann, Stefan**; **Bannert, Maria**; **Dementieva, Daryna**; **Fischer, Frank**; **Gasse, Urs**; **Groh, Georg**; **Günnemann, Stephan**; **Hüllermeier, Eyke**; **Krusche, Stephan**; **Kutyniok, Gitta**; *et al.* (2023). "ChatGPT for good? On opportunities and challenges of large language models for education". *Learning and individual differences*, v. 103, 102274.
https://doi.org/10.1016/j.lindif.2023.102274

Lin, Jialiang; **Song, Jiaxin**; **Zhou, Zhangping**; **Chen, Yidong**; **Shi, Xiaodong** (2023). "Automated scholarly paper review: Concepts, technologies and challenges". *Information fusion*, v. 98, 101830.
https://doi.org/10.1016/j.inffus.2023.101830

Lira, Rodrigo-Pessoa-Cavalcanti; **Rocha, Eduardo-Melani**; **Kara-Junior, Newton**; **Costa, Dácio-Carvalho**; **Procianoy, Fernando**; **De-Paula, Jayter-Silva**; **Gracitelli, Carolina P. B.**; **Prata, Tiago-da-Silva**; **Regatieri, Caio V.**; **Biccas-Neto, Laurentino**; **Alves, Monica** (2023). "Challenges and advantages of being a scientific journal editor in the era of ChatGPT". *Arquivos brasileiros de oftalmologia*, v. 86, n. 3, pp. 5-7.
https://doi.org/10.5935/0004-2749.2023-1003

Marcus, Gary (2022). "How come GPT can seem so brilliant one minute and so breathtakingly dumb the next?". *Marcus on AI*, December 1.
https://garymarcus.substack.com/p/how-come-gpt-can-seem-so-brilliant

Monroy, Fuensanta; **González-Geraldo, José L.** (2022). "Development of a procrastination scale in Spanish and measurement of education students' procrastination levels". *Bordón. Revista de pedagogía*, v. 74, n. 2, pp. 63-76.
https://doi.org/10.13042/Bordon.2022.93054

Peña-Fernández, Simón; **Meso-Ayerdi, Koldobika**; **Larrondo-Urena, Ainara**; **Díaz-Noci, Javier** (2023). "Sin periodistas, no hay periodismo. La dimensión social de la inteligencia artificial generativa en los medios de comunicación". *Profesional de la información*, v. *32*, n. 2, e320227.
https://doi.org/10.3145/epi.2023.mar.27

Perkins, Mike (2023). "Academic integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond". *Journal of university teaching & learning practice*, v. 20, n. 2, Article 07.
https://doi.org/10.53761/1.20.02.07

Rudolph, Jürgen; **Tan, Samson**; **Tan, Shannon** (2023). "ChatGPT: Bullshit spewer or the end of traditional assessments in higher education". *Journal of applied learning & teaching*, v. 6, n. 1.
https://doi.org/10.37074/jalt.2023.6.1.9

Santandreu-Calonge, David; **Medina-Aguerrebere, Pablo**; **Hultberg, Patrik**; **Shah, Mariam-Aman** (2023). "Can ChatGPT improve communication in hospitals?". *Profesional de la información*, v. 32, n. 2, e320219.
https://doi.org/10.3145/epi.2023.mar.19

Schulz, Robert; **Barnett, Adrian**; **Bernard, René**; **Brown, Nicholas J.L.**; **Byrne, Jennifer A.**; **Eckmann, Peter**; **Gazda, Małgorzata A.**; **Kilicoglu, Halil**; **Prager, Eric M.**; **Salholz-Hillel, Maia**; **Ter-Riet, Gerben**; **Vines, Timothy**; *et al.* (2022). "Is the future of peer review automated?". *BMC research notes*, v. 15, n. 203.
https://doi.org/10.1186/s13104-022-06080-6

Severin, Anna; **Strinzel, Michaela**; **Egger, Matthias**; **Barros, Tiago**; **Sokolov, Alexander**; **Mouatt, Julia-Vilstrup**; **Müller, Stefan** (2022). "Journal impact factor and peer review thoroughness and helpfulness: A supervised machine learning study". *arXiv*, 2207.09821.
https://doi.org/10.48550/arXiv.2207.09821

Sok, Sarin; **Heng, Kimkong** (2023). "ChatGPT for education and research: a review of benefits and risks". *Social science research network (SSRN)*, March 9.
https://doi.org/10.2139/ssrn.4378735

Srivastava, Mashrin (2023). "A day in the life of ChatGPT as an academic reviewer: Investigating the potential of large language model for scientific literature review". *OSF preprints*, February 16.
https://doi.org/10.31219/osf.io/wydct

Švab, Igor; **Klemenc-Ketiš, Zalika**; **Zupanič, Saša** (2023). "New challenges in scientific publications: Referencing, artificial intelligence and ChatGPT". *Slovenian journal of public health*, v. 62, n. 3, pp. 109-112.
https://doi.org/10.2478/sjph-2023-0015

Tlili, Ahmed; **Shehata, Boulus**; **Adakwah, Michael-Agyemang**; **Bozkurt, Aras**; **Hickey, Daniel T.**; **Huang, Ronghuai**; **Agyemang, Brighter** (2023). "What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education". *Smart learning environments*, v. 10, n. 15.
https://doi.org/10.1186/s40561-023-00237-x

Vaswani, Ashish; **Shazeer, Noam**; **Parmar, Niki**; **Uszkoreit, Jakob**; **Jones, Llion**; **Gómez, Aidan N.**; **Kaiser, Łukasz**; **Polosukhin, Illia** (2017). "Attention is all you need". In: *NIPS'17: Proceedings of the 31st international conference on neural information processing systems*, pp. 6000-6010.
https://dl.acm.org/doi/pdf/10.5555/3295222.3295349

Wang, Xuezhi; **Wei, Jason**; **Schuurmans, Dale**; **Le, Quoc**; **Chi, Ed**; **Narang, Sharan**; **Chowdhery, Aakanksha**; **Zhou, Denny** (2022). "Self-consistency improves chain of thought reasoning in language models". *arXiv*, 2203.11171v4.
https://doi.org/10.48550/arXiv.2203.11171

Zhai, Xiaoming (2023). "ChatGPT for next generation science learning". *Crossroads*, v. 29, n. 3, pp. 42-46.
https://doi.org/10.1145/3589649

7. Anexo 1

Valoraciones obtenidas. *ChatPDF* (rev. 1 y rev. 3. *GPT-3.5 turbo*) y *Bing* (rev. 2. *GPT-4*)

	Artículo 1			Artículo 2			Artículo 3			Artículo 4			Artículo 5		
	Rev. 1	Rev. 2	Rev. 3	Rev. 1	Rev. 2	Rev. 3	Rev. 1	Rev. 2	Rev. 3	Rev. 1	Rev. 2	Rev. 3	Rev. 1	Rev. 2	Rev. 3
Clasif.	B	B	B	B	B	A	A	B	A	B	B	B	A	A	A
1	1	1	-	-	5	5	1	5	1	1	5	1	2	5	1
	1	1	2	-	5	2	1	5	3	1	4	1	1	5	1
2	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
	5	-	5	5	5	5	5	5	5	5	5	5	5	3	5
3	5	4	5	5	5	5	5	5	4	5	5	5	5	4	5
4	5	1	5	5	1	5	5	5	5	5	5	5	5	2	5
	5	1	5	5	1	5	5	5	5	5	5	5	5	1	5
5	5	1	5	5	1	-	5	-	5	1	-	1	1	1	5
6	5	4	5	5	5	5	5	5	5	5	5	5	5	4	5
7	3	3	3	3	1	3	4	-	4	1	-	1	3	3	4
8	5	4	5	2	5	5	5	5	5	4	5	5	5	4	5
	5	4	5	4	5	5	5	5	5	5	5	5	5	5	5
9	5	4	4	4	5	4	5	5	5	5	5	5	5	5	5
10	5	4	5	4	5	5	5	5	5	5	5	5	5	5	5
11	4	-	4	4	-	5	4	-	4	4	-	5	4	-	4
12	5	3	4	4	4	5	4	4	5	5	4	5	5	5	5
13	3	-	5	4	-	5	5	3	4	5	5	5	5	5	5
14	1	-	3	-	-	5	4	3	4	3	-	1	4	4	4
15	1	-	3	-	-	5	4	3	4	3	3	3	5	5	5
16	1	4	3	4	-	5	4	3	4	3	5	5	5	5	5
17	4	1	4	4	4	5	4	-	4	5	5	5	4	5	4
Decisión	B	C	B	B	B	B	B	A	B	B	A	B	B	B	A

Clasificación: A) Investigación empírica (cuantitativa o cualitativa), B) Investigación teórica, ensayo, C) Experiencia o innovación educativa, y D) Otros. Criterios: 1) Formato IMRyD del resumen / Extensión del resumen, 2) Adecuación del Título / palabras clave, 3) Corrección ortográfica y sintáctica, 4) Normas *APA* / coherencia entre citas y referencias bibliográficas, 5) Tablas y figuras, 6) Interés del artículo para la comunidad educativa, 7) Generalización de los resultados, 8) Originalidad del trabajo / aportación al conocimiento educativo, 9) Introducción y justificación de la importancia del tema, 10) Fundamentación teórica, 11) Actualidad de las fuentes citadas, 12) Formulación de objetivos, 13) Proceso de recogida y análisis de información, 14) Descripción del procedimiento de muestreo, 15) Proceso de recogida y análisis de información, 16) Presentación y descripción de resultados, 17) Conclusiones y discusión.

Decisión final: A) Publicar tal y como está o con pequeñas modificaciones de redacción y/o formato, B) Podría publicarse una vez realizadas las correcciones y mejoras sugeridas, C) No publicar por los motivos especificados.