Consuegra-Fernández, M., Sanz-Aznar, J., Burguera-Serra, J. G., y Caballero, J. J. (2024). ChatGPT: el dilema sobre la autoría de las actividades evaluables en educación universitaria. *Revista de Investigación Educativa, 42*(2).
DOI: https://doi.org/10.6018/rie.565391

ChatGPT: el dilema sobre la autoría de las actividades evaluables en educación universitaria

ChatGPT: The Dilemma of the Authorship of Graded Assignments in Higher-Education

Marta Consuegra-Fernández*[1], Javier Sanz-Aznar*, Joan Gabriel Burguera-Serra* y Juan José Caballero Molina*
*Facultat de Filologia i Comunicació. Universitat de Barcelona (España)

Resumen

La irrupción de ChatGPT plantea nuevos desafíos en el ámbito educativo. Entre ellos, destaca el debate abierto en torno a las consecuencias -potencialmente negativas- que los usos de la herramienta pueden generar en los procesos de aprendizaje y evaluación del alumnado d. El siguiente trabajo explora tanto el grado de conocimiento y percepción sobre ChatGPT del profesorado universitario, como la capacidad para distinguir trabajos de autoría humana de otros originados por la inteligencia artificial. Para ello, 51 docentes de la Universidad de Barcelona, de áreas de conocimiento asociadas a la comunicación y la filología, tuvieron acceso a los textos resultantes de una actividad académica real, a partir de versiones redactadas por los propios alumnos, y de otras generadas ad hoc a través de ChatGPT. Los resultados revelaron un porcentaje de acierto en la asignación de autoría del 31 %, un valor que evidencia un nuevo obstáculo en los procesos de enseñanza, aprendizaje y evaluación en la educación superior. Paralelamente, se observó que tendían a otorgar una valoración más positiva a las muestras elaboradas por ChatGPT frente a aquellas redactadas por el alumnado. Finalmente, el artículo recoge una serie de propuestas para anticipar el impacto que podría tener un uso deshonesto de la inteligencia artificial en la adquisición de competencias y habilidades del alumnado universitario.

Palabras clave: ChatGPT; Procesamiento del lenguaje natural; comunicación escrita; educación universitaria.

1 **Correspondencia:** Marta Consuegra-Fernández, mconsuegra@ub.edu, Universitat de Barcelona. Gran Via de les Corts Catalanes, 585, 08007 Barcelona.

Abstract

The emergence of ChatGPT poses new challenges in the educational field. Among them, the open discussion on the potentially negative consequences that the program's use may generate in the learning and evaluation processes of students. The present study investigates the level of knowledge and perception of ChatGPT among university educators, as well as their proficiency in discerning student-authored texts from those generated by artificial intelligence. For this purpose, 51 professors at the University of Barcelona, specializing in communication and philology, were presented with a sample of texts extracted from an authentic academic assignment that included versions written by students themselves, together with outputs generated ad hoc by ChatGPT. The accuracy rate of the authorship assignment performed by teachers was 31%, a value that reveals a new obstacle in teaching, learning, and evaluation processes in higher education. Additionally, there was a tendency for ChatGPT-generated texts to be rated more favorably than those written by the students themselves. Finally, the article presents several suggestions aimed at anticipating the potential impact of the unethical use of artificial intelligence on the development of skills and abilities among university students.

Keywords: ChatGPT; Natural Language Processing; Written Communication; University Education.

Introducción y objetivos

Los avances en los campos de la inteligencia artificial y el aprendizaje automático (Chiche y Yitagesu, 2022; Bharadiya, 2023) han contribuido significativamente al crecimiento de la lingüística computacional. Una de las técnicas con mayor recorrido en las últimas dos décadas es el Procesamiento del Lenguaje Natural (PLN) que, a través de modelos computacionales y algoritmos, permite el análisis de textos libres y la extracción de información relevante (Locke et al., 2021)it can aid in the prediction of patient outcomes, augment hospital triage systems, and generate diagnostic models that detect early-stage chronic disease. These applications may be particularly useful in critical care where there is more patient data to analyse and prediction of patient mortality is routine. In addition to its natural language understanding (NLU. Esta utilidad ahonda tanto en la capacidad de comprensión como en la de generación del propio lenguaje natural, de manera que sus aplicaciones proporcionan una interfaz para que las personas usuarias formulen preguntas o instrucciones, y accedan a la información en forma de programas informáticos diseñados para simular, entre otras, una conversación humana y realizar una amplia variedad de tareas, los llamados *chatbots* (Luo et al., 2022).

El PLN ha exhibido un gran potencial con el lanzamiento de ChatGPT, un modelo de aprendizaje automático de OpenAI que utiliza algoritmos basados en más de 150.000 millones de características del lenguaje humano[2] y da respuesta a las instrucciones (o *prompts)* de las usuarias. ChatGPT utiliza un modelo de aprendizaje masivo a partir de textos accesibles en internet, cuya revisión automática le permite deducir y reconocer estructuras lingüísticas compartidas. Este aprendizaje se actualiza a medida que el

2 Disponible en https://openai.com/blog/chatgpt/

programa accede y compara nuevos textos con patrones previamente aprehendidos, perfeccionando así su habilidad de PLN. Actualmente el programa dispone de distintas versiones, entre las que se encuentran GPT-3.5, lanzado en noviembre de 2022, correspondiente al prototipo de prueba y de acceso gratuito, y el modelo GPT-4 lanzado en marzo de 2023, solo accesible bajo suscripción de pago[3].

Si bien el PLN viene desarrollándose desde finales del s. XX (Manning y Schutze, 1999), la tecnología GPT (del inglés, *Generative Pre-training Transformer*) (Zhu y Luo, 2022) supone un revolucionario avance cualitativo y ha sido considerada por algunos autores y autoras como una innovación *disruptiva* (Dowling y Lucey, 2023)using a runtime and mobile services has gotten more complex along with the composition of a large number of atomic services. Different services are provided by mobile cloud components to represent the non- functional properties as Quality of Service (QoS. Su uso se ha popularizado hasta erigirse en un recurso de alcance masivo. Prueba de ello es la cobertura mediática creciente que ha sido objeto desde su lanzamiento, hasta el punto de llegar a ocupar un papel destacado en la agenda pública (Fig. 1).

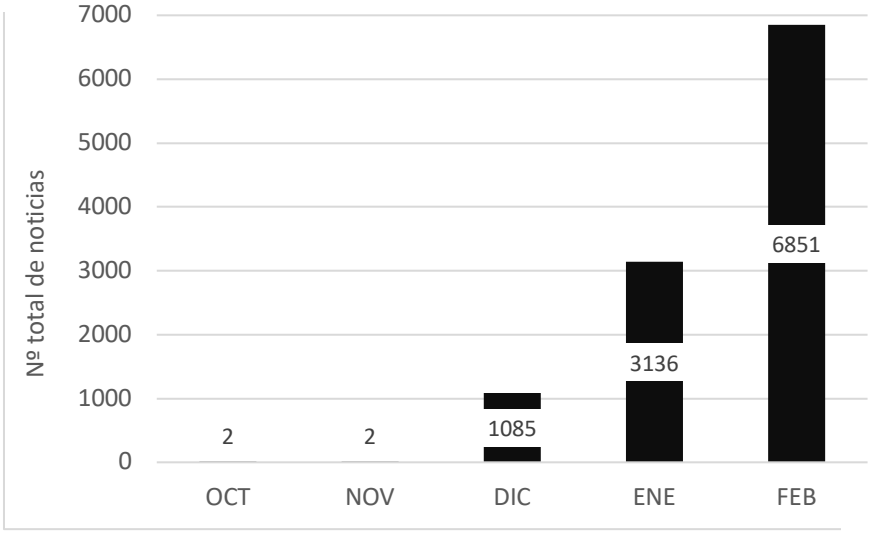

Figura 1. Noticias obtenidas con la búsqueda por palabras claves "ChatGPT" en la hemeroteca de prensa contemporánea de España (Fuente: *MyNews*).

Esta notoriedad mediática ha despertado el interés de usuarios procedentes de múltiples áreas (Rudolph et al., 2023) y, en especial, de la comunidad de estudiantes (Haque et al., 2022).

3 Disponible en https://openai.com/blog/chatgpt/

En el sector educativo, las personas usuarias recurren a ChatGPT para diversas tareas que comprenden la composición textual o la obtención y el procesamiento de información. Sin embargo, a diferencia de sectores como el marketing digital, donde el uso de ChatGPT está más aceptado, su uso con fines educativos resulta controvertido (Baidoo-Anu y Owusu, 2023). Mientras algunos perciben los sistemas de PLN como una posible herramienta didáctica eficiente, accesible y capaz de reducir costos de las instituciones educativas (Heller et al., 2005; Pérez et al., 2020; Tallyn et al., 2018; Villegas-Ch et al., 2020), otros consideran que no puede sustituir la interacción humana, fundamental para el desarrollo social y comunicativo de del alumnado (Butnaru et al., 2021; Sharp y Theiler, 2018). Por otra parte, ChatGPT ha recibido también críticas por la potencial presencia de sesgos ideológicos en el algoritmo que, como ha sucedido en otras ocasiones, supondría una amenaza para el bienestar de algunos colectivos (Doshi et al., 2023). Con todo, uno de los temas más controvertidos respecto a ChatGPT radica en la eventual apropiación indebida, en términos de autoría, que los estudiantes pueden hacer de los textos y contenidos producidos por la herramienta. De ser así, no solo se cuestionaría el propio proceso de evaluación por parte de los docentes (Baidoo-Anu y Owusu, 2023), sino que se verían seriamente comprometidos el aprendizaje y la capacidad de pensamiento crítico que debe acreditar el alumnado (García-Peñalvo, 2023). En este punto cabe destacar también la tendencia, cada vez mayor, de tareas académicas y evaluaciones no presenciales (Gómez y Alende, 2022; Sáiz-Manzanares et al., 2022), que inevitablemente conllevan un menor control de autoría por parte del docente, así como un mayor riesgo de plagio y fraude académico. Prueba de ello es la popularidad de las herramientas de detección de plagio, cada vez más extendidas en los procesos de preevaluación universitarios (Khalil y Er, 2023).

Este temor tiene su base en evidencias recogidas por distintos usuarios que han sometido ChatGPT a distintas pruebas evaluativas. Entre otras, parece que este programa es capaz de aprobar exámenes de Derecho (Bommarito y Katz, 2022; Choi et al., 2023) u obtener la licencia médica para ejercer en Estados Unidos (Kung et al., 2023). En España, también ha sido capaz de superar la nota de corte del examen MIR de 2022 (Carrasco et al., 2023). Sin embargo, hasta la fecha de elaboración de este manuscrito, solo se ha publicado un artículo *pre-print* que haya explorado, con una metodología científica, los riesgos reales de un posible uso ilegítimo en el ámbito científico-académico (Else, 2023). Un grupo de la Escuela de Medicina Feinberg de la Universidad Northwestern en Estados Unidos (Gao et al., 2022) ha evidenciado recientemente que ChatGPT es capaz de redactar, a partir de contenido inventado, resúmenes científicos verosímiles. Concretamente, tras una lectura detenida, un grupo de investigadores, previamente advertidos del objetivo del estudio, fueron capaces de asignar correctamente la autoría artificial de los textos en un 68 % de los casos (Gao et al., 2022).

Este escenario determina la necesidad de evaluar su posible impacto, así como la pertinencia de articular medidas preventivas en caso de ser necesario. Si bien es cierto que han ido apareciendo sistemas computacionales que permiten detectar si un texto en inglés ha sido generado mediante PLN (tales como GPTZero o Contentatscale AI Content Detector), no hay evidencia de la eficacia de estas herramientas en español.

En este contexto, la hipótesis de la presente investigación defiende que el profesorado universitario no dispone de mecanismos efectivos para distinguir textos elaborados por estudiantes de aquellos generados por inteligencia artificial. De validarse esta hipótesis, el impacto en términos de riesgo derivado de la irrupción de ChatGPT en la educación superior pondría en cuestión los procesos de evaluación de cierto tipo de actividades académicas.

Los objetivos de la presente investigación son:

- Examinar la valoración que, en términos cualitativos generales, atribuyen los docentes universitarios a textos anonimizados elaborados bien por estudiantes, bien por ChatGPT.
- Evaluar el nivel de acierto de del profesorado en la asignación de la autoría (estudiantes vs. ChatGPT) de dichos textos.
- Analizar el grado de conocimiento, percepción del impacto y necesidades del profesorado frente al uso de la inteligencia artificial en un contexto universitario.

Método

La investigación se ha realizado en colaboración con una muestra de docentes en activo de las distintas titulaciones de la Facultad de Filología y Comunicación (FiC) de la Universidad de Barcelona (UB) y responde a la necesidad de dotar de evidencias empíricas a algunas de las reflexiones que se están sosteniendo en relación con la incursión de ChatGPT en el sector educativo. Se trata de un estudio de diseño transversal, descriptivo y analítico que combina la metodología cuantitativa y la cualitativa, y utiliza la técnica del cuestionario con ítems adaptados de referencias previas. Brevemente, se pidió a docentes con formación y habilidades pedagógicas específicas en lengua, análisis literario y comunicación que, tras leer tres textos anonimizados correspondientes a una actividad evaluativa real, los jerarquizaran cualitativamente y, con posterioridad, señalaran si los textos objeto de análisis habían sido producidos por estudiantes o por ChatGPT.

Población y Muestra

Para la presente investigación se obtuvo una muestra de 51 docentes de la FiC de la UB que formaban parte del Personal Docente e Investigador durante el curso 2022-23. el profesorado está adscrito a los departamentos de Filología Clásica, Románica y Semítica; Filología Hispánica, Teoría de la Literatura y Comunicación; Filología Catalana y Lingüística General; y Lenguas y Literaturas Modernas y Estudios Ingleses. Por su perfil académico y profesional, acreditan un amplio conocimiento de la lengua, la comunicación y la cultura, y un gran dominio del análisis e interpretación de distintas tipologías de textos. Su formación y dedicación tanto investigadora como docente les confiere la consideración de especialistas en cuanto al análisis de textos.

La muestra (Tabla 1) está constituida por docentes procedentes de Comunicación, Filologías, Estudios Literarios y/o Lingüística, de edades comprendidas entre 32 y 70 años, y con una experiencia docente heterogénea que tiene su promedio en los 21 años.

Tabla 1

Descripción de la muestra de docentes (N=51)

Variables	Valores descriptivos
Edad	*(años)*
1. Promedio	50,39
2. Desviación estándar	±9,37
3. Moda	47
4. Rango de edad	[32,71]
Género	*(individuos)*
1. Femenino	27 (59,94%)
2. Masculino	22 (43,14%)
3. Sin determinar	2 (3,92%)
Área de conocimiento	*(individuos)*
1. Filologías, estudios literarios y/o lingüística	42 (82,35%)
2. Comunicación audiovisual y publicidad	4 (7,84%)
3. Periodismo	5 (9,80%)
Años dedicados a la docencia universitaria	*(años)*
1. Promedio	21,06
2. Desviación	±10,28
3. Moda	15
4. Rango de experiencia	[4,49]

Instrumento

Para la realización del experimento se eligió la reseña literaria como género textual objeto de análisis por parte del equipo docente, y se obtuvieron 6 reseñas en total, 3 de ellas elaboradas por ChatGPT y 3 realizadas por alumnos.

Como modelo de lenguaje entrenado a través de miles de textos públicos en internet, ChatGPT es capaz de dar respuesta a múltiples preguntas de distinta índole y también de elaborar textos académicos de distintas tipologías[4]. No obstante, con el objetivo de evitar la influencia de variaciones inherentes al patrón y estilo del ejercicio, solo se incluyeron reseñas literarias. La reseña cumple con los requerimientos específicos de una entrega evaluable presente en los planes docentes de todas las titulaciones enmarcadas en FiC, por lo que los docentes participantes en la investigación están familiarizados con su lectura y evaluación. Asimismo, se trata de una tarea de elaboración compleja que busca conocer la capacidad del alumno a la hora de llevar a cabo un análisis crítico y objetivo de un producto cultural.

Las reseñas escritas por alumnos integradas en el experimento provienen de muestras reales desarrolladas en el curso 2021-2022, en el marco de la asignatura "Géneros y

4 Disponible en https://openai.com/blog/chatgpt/

formatos de la comunicación escrita" del Grado de Comunicación e Industrias Culturales de FiC. Se trata de una asignatura obligatoria de segundo curso. Se eligieron muestras del curso 21-22, por ser anteriores a la existencia de ChatGPT, a fin de evitar que las reseñas ya pudieran haber estado total o parcialmente desarrolladas mediante ChatGPT. Además, se definió una entrega concreta de una única asignatura para asegurar que cada texto procediera de un alumno o alumna diferente, garantizando así la heterogeneidad de la muestra.

Tras anonimizar los textos recopilados, 40 en total, la selección se hizo de acuerdo con los siguientes criterios de inclusión: (*i*) textos que versaran, cada uno de ellos, sobre obras diferentes, y (*ii*) textos que recibieran una valoración mínima de "notable". En la selección también se descartaron (*i*) textos con una extensión superior a 600 palabras, (*ii*) textos con subtítulos, imágenes, gráficos u otros elementos estructurales o de diseño propios, (*iii*) textos con una división estructural inferior a 3 párrafos, (*iv*) textos redactados en primera persona del singular, y (*v*) textos con un porcentaje de plagio superior al 1 % según el sistema Urkund. De las 7 reseñas resultantes, se seleccionaron de forma aleatoria 3 de ellas (disponibles en el Apéndice).

Los 3 textos elaborados por ChatGPT en su versión GPT-3 se obtuvieron a través de la plataforma de OpenAI en respuesta a la siguiente instrucción: "*Escribe una reseña literaria en castellano de 500 palabras de extensión sobre el libro...*" (consultar el Apéndice). En los tres casos se concretó el título de la obra objeto de reseña que, para ser concordante con los libros elegidos por el alumnado, correspondían también a obras de narrativa periodística (Tabla 2). La instrucción facilitada a ChatGPT es idéntica al enunciado de la actividad evaluadora que realizaron los alumnos con el objeto de fijar unas mismas condiciones basales de ejecución. Si bien es cierto que la plataforma permite incluir más instrucciones que añadan complejidad o calidad al texto resultante, la elección de una instrucción básica evita la influencia de la experiencia y del conocimiento de usuario que opera ChatGPT, rasgos que podrían actuar como variables de confusión.

Tabla 2

Obras literarias sobre las que narran las tres reseñas incluidas en la investigación y número de docentes que revisaron cada una de ellas.

Obra	**Autor**	**Año de edición**	**Autoría de la reseña**	**Nº de docentes expuestos a cada muestra**
El hijo del chófer	Jordi Amat i Fusté	2020	ChatGPT	25
A sangre fría	Truman Capote	1965	ChatGPT	23
Inshallah	Orianda Fallaci	1992	ChatGPT	28
She said	Jodi Kantor y Megan Twohey	2022	Alumno	25
El periodista y el asesino	Janet Malcom	2004	Alumno	28
Ébano	Ryszard Kapuściński	2006	Alumno	24

Con el fin de explorar la percepción del profesorado respecto a las reseñas, se elaboró un cuestionario (disponible en el Apéndice) con Pavlovia[5], una plataforma de acceso abierto que ofrece herramientas para crear, ejecutar y recopilar datos en línea. Las preguntas se estructuraron en cuatro bloques ordenados de la siguiente manera: (*i*) lectura y prelación de las reseñas a partir de una valoración cualitativa general entendida en sentido amplio (calidad de contenidos, aspectos formales, etc.), (*ii*) atribución de la autoría de cada texto a un alumno o a la inteligencia artificial (se contempló como respuesta la imposibilidad de definir la autoría), (*iii*) caracterización sociodemográfica y formativa de cada participante y, (*iv*) conocimiento y autopercepción, evaluada mediante una escala Likert discreta de 6 grados desde 1 (totalmente en desacuerdo) hasta 6 (totalmente de acuerdo), relativa al conocimiento que el profesorado tenía de la herramienta ChatGPT, su opinión respecto a su utilidad en el sector educativo, la desconfianza y percepción de amenaza, la necesidad de recibir formación específica sobre ella y la previsión de su alcance futuro. Para simular un contexto evaluativo real y evitar sesgos o condicionamientos previos en el profesorado en la asignación de autorías, no se hizo referencia alguna al propósito de la investigación. Por otra parte, el profesorado solo obtuvo acceso a las preguntas relativas a su conocimiento y percepción de ChatGPT una vez terminadas las tareas de evaluación cualitativa y asignación de autoría de los textos, sin posibilitar retrocesos que permitieran repetir el proceso y corregir su valoración inicial.

Los 9 ítems evaluados en la escala Likert se adaptaron de la publicación de Escoda y Conde (2016) y, para garantizar su fiabilidad, se calcularon los valores de alfa de Cronbach correspondientes a cada ítem. En todos los casos se obtuvieron valores iguales o superiores a 0,7, umbral de fiabilidad establecido para muestras pequeñas (Bujang et al., 2018).

Previamente a su distribución, se comprobó el correcto funcionamiento de la herramienta mediante una prueba piloto realizada a 4 sujetos que cumplían los mismos requisitos que los participantes de la muestra y que no formaron parte del estudio.

Procedimiento de recogida y análisis de datos

Cada docente analizó tres reseñas que se asignaron mediante un proceso pseudoaleatorio para evitar sesgos por combinatoria, garantizando siempre que (*i*) cada participante leyera textos diferentes, (*ii*) que de esos textos, por lo menos, uno o dos hubieran sido elaborados por ChatGPT, y uno o dos por alumnos, (*iii*) que todos los textos se leyeran y revisaran un número de veces similar, (*iv*) que todos los textos se mostraran el mismo número de veces en primer, segundo o tercer lugar y, finalmente, (*v*) que todas las situaciones en las que se presentaran dos textos de una autoría fueran las mismas que en las que se presentaban dos textos de la otra.

En todos los casos se informó a los sujetos que su participación era anónima y voluntaria, y que todos los datos recabados serían utilizados únicamente por la UB, excluyendo la cesión a terceros. Asimismo, se les informó acerca del compromiso de cumplir con lo que establece el Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril, relativo a la protección de las personas físicas en lo que respecta al tratamiento de sus datos personales y a la libre circulación de estos datos y

5 Disponible en https://pavlovia.org/

la Ley Orgánica 3/2018, de 5 de diciembre, de protección de datos personales y garantía de los derechos digitales.

La selección de los y las participantes se hizo mediante un muestreo no probabilístico por conveniencia. El cuestionario se envió, el 25 de febrero de 2023 a través de una cuenta de correo electrónico de la UB propia del proyecto, a 163 docentes de la FiC de la UB de un total de 358[6], que formaban parte del Personal Docente e Investigador durante el curso 2022-23. Se obtuvo un índice de participación del 31,29 %, equivalente a 51 cuestionarios cumplimentados que constituyeron la muestra final representativa (con un nivel de confianza del 95 % y un margen de error del 13 %).

Las respuestas se recopilaron y analizaron mediante IBM SPSS *Statistics* v.26. Los resultados numéricos de la asignación y valoración de los textos se exponen en números absolutos y en porcentajes de frecuencias relativas. Para analizar las asociaciones entre las dos variables nominales "autoría ChatGPT" y "autoría alumnos" se calcularon, a partir de las frecuencias observadas y esperadas, los componentes de Chi-cuadrado de Pearson y los índices de Cramer para cuantificar entre 0 y 1 el grado de relación. La significación estadística se estableció a partir de $p<0,05$ y el nivel de confianza en 95 %.

Por su parte, las respuestas del cuestionario evaluadas con escala Likert se representan con la media (M), la mediana (Md) y el error estándar de la media (SEM) del total de respuestas. Adicionalmente, también se analizaron los ítems de naturaleza cualitativa en los que se solicitaba una respuesta redactada.

Resultados y discusión

Las reseñas fueron revisadas un total de 153 veces por los docentes (Figura 2). Tras estas lecturas, en 48 casos (31,37 %) la autoría se asignó a alumnos y en 43 ocasiones a ChatGPT (28,10 %), mientras que hasta en 62 casos el profesorado alegó "no poder identificar la autoría" (40,52 %).

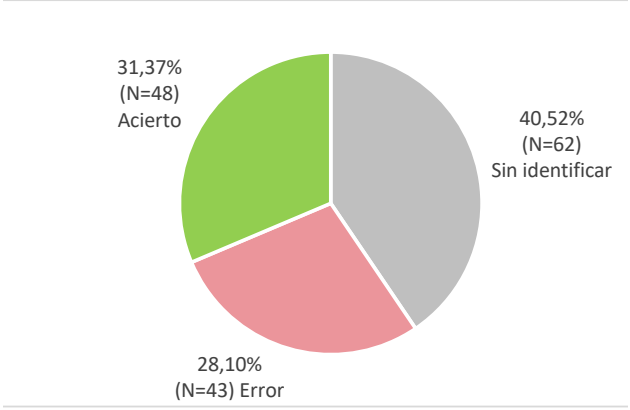

Figura 2. Porcentaje y números absolutos de asignaciones correctas (acierto), incorrectas (error) y no identificadas en relación con la autoría de los textos.

6 Extraído el 21 de febrero de 2023 de la web de la Facultat de Filologia i Comunicació de la Universitat de Barcelona https://www.ub.edu/portal/web/filologia-comunicacio

Como muestra la Figura 2, el porcentaje de acierto de estas asignaciones fue del 31,37 %, un valor inferior al 68 % de aciertos obtenidos por el equipo investigador que analizaron resúmenes científicos en el estudio de Gao et al. (2022) with unknown information on the accuracy and integrity of using these models in scientific writing.\n\nMethods We gathered ten research abstracts from five high impact factor medical journals (n=50. Esta primera constatación, junto con el elevado número de reseñas sin identificar, revela el riesgo real de que el posible fraude pase inadvertido. La diferencia con respecto al estudio de Gao et al. (2022) with unknown information on the accuracy and integrity of using these models in scientific writing.\n\nMethods We gathered ten research abstracts from five high impact factor medical journals (n=50 se debe a que, en su caso, el equipo investigador informó previamente a los revisores de la existencia de resúmenes generados por ChatGPT, mientras que, en el presente estudio, el profesorado desconocía el propósito de la lectura de los textos. Asimismo, tampoco se permitió que retrocedieran en el cuestionario para llevar a cabo una lectura más detallada y una búsqueda específica de indicios de escritura con sistemas PLN. Esta diferencia metodológica entre los dos estudios resulta fundamental para entender el porcentaje de identificación de la autoría de los textos revisados por los docentes en nuestra investigación, y evidencia la necesidad de que el profesorado conozca el programa y las consecuencias que de su uso podrían derivarse.

Tabla 3

Análisis de los aciertos, errores y no identificados de autorías asignadas a alumnos y a ChatGPT

	% Acierto (N)	**% Error (N)**	**% S/iden. (N)**	**Chi cuadrado Pearson**	**V de Cramér**	**Sig. bilater.**
ChatGPT	13,1% (20)	13,7% (21)	22,9% (35)	2,384	0,125	0,304
Alumnos	18,3% (28)	14,4% (22)	17,6% (27)			

Además, el análisis en profundidad de los datos reveló que los aciertos, los errores y las asignaciones no identificadas se distribuían uniformemente tanto si se trataba de reseñas realizadas por alumnos, como de aquellas elaboradas por ChatGPT (Tabla 3), lo cual sugiere que no había rasgos propios de la redacción de una autoría u otra que permitiera distinguirlos. No obstante, tras preguntar por los motivos por los que habían vinculado ciertas reseñas a la inteligencia artificial, el 60,8 % respondió alegando motivos que se han agrupado por temáticas coincidentes en la Tabla 4, frente al resto que no contestó o esgrimió su desconocimiento.

Tabla 4

Motivos de asignación de textos a ChatGPT agrupados por temáticas

Temáticas	Respuestas de los participantes
Estructura	"La estructura, demasiado mecanizada. Aunque en los tres textos se da una estructura (problemática) similar."
Calidad escrita	"... en comparación con los otros dos, muestra una expresión escrita más pobre."
Reiteraciones	"... es esencialmente reiterativo y cada párrafo está construido de forma muy semejante desde el punto de vista sintáctico."
Errores de puntuación	"...la puntuación y la complejidad de las frases me generan esta duda." "...algunas faltas de concordancia y comas"
Modalización enunciativa	"El carácter neutro e impersonal del estilo", "La expresión de uno de los textos me ha parecido meramente informativa, construida con recursos estilísticos tipificados, sin mucha personalidad."
Traducciones literales	"La secuencia 'la familia Catalana' se usa (…) con mayúsculas iniciales, lo cual es sospechoso de que parte de un original en otra lengua, posiblemente el inglés. ", "...usos léxicos más propios de obras traducidas."

Entre los argumentos que el grupo docente alegó como elementos discriminatorios entre los textos del alumnado y los de la inteligencia artificial, se encontró la presencia de reiteraciones literales y estructuras equivalentes parafraseadas como aspecto más frecuente (20 %). Esta es una de las características más notable del estilo de redactado propio de ChatGPT que puede llegar incluso a evadir sistemas de detección de plagio como GPTZero, GPT-2 Output Detector o AI Detector (Anderson et al., 2023). En menor frecuencia, el grupo docente argumentó la presencia de una modalización enunciativa neutral y poco expresiva, errores de coherencia discursiva (puntuación, secuenciación textual, etc.) y una aparente disponibilidad léxica incongruente con la presumible en el alumnado. Además, la estructura textual mecanizada, uno de los elementos distintivos de la redacción de ChatGPT, ya observada en publicaciones previas (Gao et al., 2022), se mencionó en dos ocasiones. En general, se observó que los criterios del profesorado coincide entre ellos como también lo hacen con los recogidos en la escasa literatura disponible hasta la fecha (Gao et al., 2022). No obstante, no siempre resultaron suficientemente concluyentes como para determinar correctamente la autoría y, en ocasiones, incluso los condujeron a una identificación errónea.

Tras observar que existe un riesgo real de incorrecta identificación de la autoría de las reseñas, se procedió a analizar la prelación cualitativa que el profesorado había realizado de los textos.

Tabla 5

Prelación de las reseñas a partir de la valoración global cualitativa

	% 1º (N)	**% 2º (N)**	**% 3º (N)**	**Chi cuadrado Pearson**	**V de Cramér**	**Sig. bi-later.**
ChatGPT	18,3% (28)	19,0% (29)	12,4% (19)			
				4,758	0,176	0,09
Alumnos	15,0% (28)	14,4% (22)	20,9% (32)			

Los resultados (Tabla 5) reflejaron una tendencia de priorización de textos escritos por ChatGPT frente a los textos redactados por el alumnado. No obstante, la muestra de estudio puede no ser suficiente para detectar un efecto pequeño (V Cramer=0,176), tal y como apunta el valor de significancia resultante (p=0,09). Así, pese a que puede ser necesario desarrollar una investigación futura con una muestra mayor que facilite analizar la significancia del sesgo, e incluso que incluya valoraciones numéricas, no puede obviarse la tendencia observada por el potencial impacto negativo que puede tener. En conjunto, se concluye que no solo existe un riesgo en la asignación de la autoría, sino que los textos generados por ChatGPT podrían obtener una valoración superior a los realizados por los alumnos. De ser así, el uso fraudulento de la herramienta supondría una ventaja evaluativa y podría promover dichas prácticas entre el alumnado (Agud, 2014; Díez-Martínez, 2015).

No es la primera vez que la innovación tecnológica incide sobre el ámbito educativo y cuestiona la metodología didáctica y de aprendizaje tradicional. Sin embargo, los currículos y los planes docentes actuales integran las competencias en tecnologías de la información y la comunicación (TIC) (Domingo-Coscollola et al., 2020). Esto atañe no solamente al dominio digital del alumnado, sino también la llamada Competencia Digital Docente (CDD) (Hall et al., 2014) que abarca las destrezas y conocimientos docentes para usar dichas metodologías e instrumentos como recursos didácticos en el aula. En este sentido, es importante conocer la opinión y el conocimiento docente acerca de los sistemas PLN que como ChatGPT podrían formar parte de esta transformación tecnológica. Además, el conocimiento de la herramienta pone en alerta al profesorado ante sus posibles aplicaciones ilícitas y les anticipa para reconocer autorías fraudulentas (Gao et al., 2022).

En relación con este aspecto, el análisis del cuestionario autoperceptivo respondido por el profesorado reveló que actualmente estos tienen un escaso conocimiento práctico de la herramienta (Tabla 6, *Ítem 2*) debido, probablemente, a su reciente aparición, y que expresan una clara voluntad de recibir una formación específica al respecto (Tabla 6,

Ítem 7). En línea con estos resultados, otros estudios apuntan a que la práctica educativa universitaria no está en sintonía con la realidad digital actual (Domingo-Coscollola et al., 2020; Sancho-Gil et al., 2017) e inciden en la necesidad de enriquecer la formación en CDD del profesorado para que este pueda desarrollar de forma adecuada sus labores docentes (Cervera et al., 2016; Cuartero et al., 2016).

Por otro lado, las personas participantes tampoco tienen una opinión clara acerca de ChatGPT, ni sobre si este representa una amenaza para la docencia universitaria. La tendencia observada es que el profesorado considera que la herramienta tendrá un mayor potencial en el futuro (Tabla 6, *Ítem 8*), pese a que la literatura más reciente evidencia que el uso predominante actual pertenece a la comunidad de estudiantes (Haque et al., 2022).

Tabla 6

Conocimiento y percepción evaluados mediante preguntas de repuesta Likert desde 1 (totalmente desacuerdo) a 6 (totalmente de acuerdo) a los docentes participantes (N=51).

***Ítem* de cada pregunta**	**Media (M)**	**Mediana (Md)**	**Error media (SEM)**
Ítem 1. Te muestras receptivo al uso de programas como ChatGPT en tu actividad académica	2,82	3	1,57
Ítem 2. Has utilizado ChatGPT	1,69	1	1,42
Ítem 3. Crees que estos programas suponen una amenaza para la docencia universitaria	4,04	4	1,60
Ítem 4. Crees es urge modificar la evaluación en tus asignaturas	3,92	5	1,94
Ítem 5. Crees que deben tomarse medidas al respecto	4,31	5	1,66
Ítem 6. Crees que puede ser una herramienta útil para la formación universitaria	3,37	4	1,56
Ítem 7. Crees que el profesorado universitario debería recibir formación específica sobre estas nuevas herramientas	4,76	6	1,69
Ítem 8. Crees que los estudiantes usaran este tipo de herramientas profesionalmente	4,90	5	1,33
Ítem 9. Crees que tus estudiantes utilizan este tipo de recursos en sus trabajos académicos	3,33	3	1,60

Cabe destacar que no se han identificado diferencias significativas asociadas a edad, género, formación y años de docencia (análisis mediante prueba *t de Student* para variables paramétricas y *Mann-Witney* para no paramétricas, en cada caso). Apa-

rentemente cualquier categorización con las variables mencionadas se comporta de la misma manera que los resultados globales observados en cuanto a identificación de la autoría, prelación cualitativa de los textos, y conocimiento y autopercepción medidos a través de la escala Likert. La presente investigación se ha centrado en el estudio de una muestra muy homogénea de docentes con formación y habilidades pedagógicas específicas en lengua, comunicación y análisis literario. No obstante, no se descarta la existencia de posibles diferencias entre grupos de edad, género, formación y años de docencia de las personas participantes, que podrían estudiarse si se ampliara y diversificara la muestra, incluyendo profesorado s de otras universidades y, sobre todo, de otras áreas de conocimiento.

Conclusiones

Este trabajo presenta una de las primeras evaluaciones del riesgo asociado a un uso ilegítimo de la tecnología ChatGPT y cuestiona el proceso de evaluación de ciertas tareas, a la par que sugiere la existencia de una interferencia en el aprendizaje del alumnado. La investigación efectuada corrobora la hipótesis de la investigación: las evidencias obtenidas atestiguan que el profesorado universitario de las áreas de conocimiento tratadas no dispone de mecanismos para discernir entre textos generados por ChatGPT y textos redactados por estudiantes. Además, en líneas generales, el profesorado desconoce aún este tipo de herramientas e ignora su potencial uso práctico.

Sin embargo, la irrupción de los sistemas de PLN, como ChatGPT, en el sector educativo es una realidad inevitable que presumiblemente se irá consolidando con el tiempo, como ha sucedido con otras innovaciones tecnologías que también en su momento fueron percibidas como *disruptivas*. En este sentido, no parece factible ni tampoco positivo prohibir categóricamente el uso de inteligencia artificial con fines pedagógicos y, en su lugar, deberían explorarse los mecanismos de integración en el marco formativo, como un nuevo recurso, para evaluar sus beneficios en la misma medida en la que se anticipan también sus potenciales limitaciones y perjuicios. Asimismo, parece pertinente proponer algunas actuaciones, a saber, (*i*) informar a los docentes y a los alumnos acerca de la existencia de las distintas herramientas que usan PLN y de aplicaciones derivadas de la inteligencia artificial para que conozcan tanto sus posibles utilidades como sus inconvenientes, (*ii*) formar al profesorado y al alumnado para que hagan un uso correcto y ético de dichos recursos, y eviten desviaciones que interfieran en el desarrollo del aprendizaje, (*iii*) integrar las nuevas tecnologías y herramientas en los planes docentes, facilitando guías y recursos para normalizar su implementación a nivel teórico y práctico, (*iv*) limitar su uso en los casos en los que tales herramientas condicionen severamente la adquisición de las habilidades o competencias objeto de aprendizaje-evaluación, hecho que podría exigir la presencialidad y la escritura manual en ciertos casos, y (*v*) replantear tareas o actividades de evaluación que hayan podido quedar obsoletas en el escenario de transformación tecnológica en el que nos encontramos actualmente, y sustituirlas por otras que incluyan los nuevos desafíos digitales.

La metodología presenta algunas limitaciones que deben ser consideradas. En primer lugar, el estudio incluye exclusivamente textos realizados por el programa ChatGPT (versión GPT-3), por lo que las implicaciones pueden no ser representativas

de sus actualizaciones o de otros sistemas de PLN. En segundo lugar, los resultados se obtienen del estudio concreto en torno al género 'reseña literaria' y podrían diferir de otros géneros discursivos. En tercer lugar, la calidad de los textos generados por ChatGPT depende en gran medida de la claridad y concreción de la información proporcionada en la instrucción del usuario. En este sentido, instrucciones más complejas hubiesen podido dar lugar a reseñas diferentes que, previsiblemente, habrían podido alterar la opinión del profesorado

Una posible línea de análisis futura podría reproducir una investigación similar tras ofrecer una formación específica a los equipos docentes acerca del funcionamiento de la herramienta ChatGPT y su sistema de PLN para contrastar el previsible aumento de la identificación de la autoría de los textos analizados.

Agradecimientos

Este trabajo se ha elaborado en el marco de las investigaciones promovidas por el Grup d'Innovació en Comunicació i Emprenedoria de la Universitat de Barcelona. Queremos agradecer la colaboración de los docentes de la FiC de la UB por su participación en el estudio y el interés mostrado en torno a la investigación.

Referencias

Agud, J. L. (2014). Fraude y plagio en la carrera y en la profesión. *Revista Clínica Española*, *214*(7), 410–414. https://doi.org/10.1016/J.RCE.2014.03.007

Anderson, N., Belavy, D. L., Perle, S. M., Hendricks, S., Hespanhol, L., Verhagen, E., y Memon, A. R. (2023). AI did not write this manuscript, or did it? Can we trick the AI text detector into generated texts? The potential future of ChatGPT and AI in Sports & Exercise Medicine manuscript generation. *BMJ Open Sp Ex Med*, *9*, 1568. https://doi.org/10.1136/bmjsem-2023-001568

Baidoo-Anu, D., y Owusu, L. (2023). Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. *SSRN*. https://doi.org/10.2139/SSRN.4337484

Bharadiya, J. (2023). A Comprehensive Survey of Deep Learning Techniques *Natural Language Processing*. *European Journal of Technology, 7*(1), 58-66. https://doi.org/10.47672/ejt.1473

Bommarito, M. J., y Katz, D. M. (2022). GPT takes the Bar Exam. *SSRN*. http://dx.doi.org/10.2139/ssrn.4314839

Bujang, M. A., Omar, E. D., y Baharum, N. A. (2018). A Review on Sample Size Determination for Cronbach's Alpha Test: A Simple Guide for Researchers. *The Malaysian Journal of Medical Sciences : MJMS*, *25*(6), 85. https://doi.org/10.21315/MJMS2018.25.6.9

Butnaru, G. I., Haller, A. P., Dragolea, L. L., Anichiti, A., y Hârșan, G. D. T. (2021). Students' Wellbeing during Transition from Onsite to Online Education: Are There Risks Arising from Social Isolation? *International Journal of Environmental Research and Public Health*, *18*(18). https://doi.org/10.3390/IJERPH18189665

Carrasco, J. P., García, E., Sánchez, D. A., Porter, E., De La Puente, L., Navarro, J., y Cerame, A. (2023). ¿Es capaz "ChatGPT" de aprobar el examen MIR de 2022?

Implicaciones de la inteligencia artificial en la educación médica en España. *Revista Española de Educación Médica*, *4*(1). https://doi.org/10.6018/edumed.556511

Cervera, M. G., Martínez, J. G., y Mon, F. M. E. (2016). Competencia digital y competencia digital docente: una panorámica sobre el estado de la cuestión. *RiiTE Revista Interuniversitaria de Investigación En Tecnología Educativa*, 2529–9638. https://doi.org/10.6018/RIITE2016/257631

Chiche, A., y Yitagesu, B. (2022). Part of speech tagging: a systematic review of deep learning and machine learning approaches. *Journal of Big Data, 9*(1), 1-25. https://doi.org/10.1186/s40537-022-00561-y

Choi, J. H., Hickman, K. E., Monahan, A. B., y Schwarcz, D. (2023). ChatGPT goes to Law School. *SSRN*. https://dx.doi.org/10.2139/ssrn.4335905

Cuartero, M. D., Gutiérrez, I., Paz, M., Espinosa, P., y Clave, P. (2016). Análisis Conceptual de Modelos de Competencia Digital del Profesorado Universitario / Conceptual analysis of digital competence models of university teacher. *Revista Latinoamericana de Tecnología Educativa - RELATEC*, *15*(1), 97–114. https://doi.org/10.17398/1695-288X.15.1.97

Díez-Martínez, E. (2015). Deshonestidad académica de alumnos y profesores. Su contribución en la desvinculación moral y corrupción social. *Sinéctica*, *44*, 14. https://dialnet.unirioja.es/servlet/articulo?codigo=8239864&info=resumen&idioma=ENG

Domingo-Coscollola, M., Bosco, A., Segovia, S. C., y Valero, J. A. S. (2020). Fomentando la competencia digital docente en la universidad: Percepción de estudiantes y docentes. *Revista de Investigación Educativa*, *38*(1), 167–182. https://doi.org/10.6018/RIE.340551

Doshi, R. H., Bajaj, S. S., y Krumholz, H. M. (2023). ChatGPT: Temptations of Progress. *The American Journal of Bioethics*. https://doi.org/10.1080/15265161.2023.2180110

Dowling, M., y Lucey, B. (2023). ChatGPT for (Finance) research: The Bananarama Conjecture. *Finance Research Letters*, *53*, 103662. https://doi.org/10.1016/J.FRL.2023.103662

Else, H. (2023). Abstracts written by ChatGPT fool scientists. *Nature*, *613*(7944), 423. https://doi.org/10.1038/D41586-023-00056-7

Escoda, A. P., y Conde, M. J. R. (2016). Evaluación de las competencias digitales autopercibidas del profesorado de Educación Primaria en Castilla y León (España). *Revista de Investigación Educativa*, *34*(2), 399–415. https://doi.org/10.6018/RIE.34.2.215121

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., y Pearson, A. T. (2022). Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *BioRxiv*, 2022.12.23.521610. https://doi.org/10.1101/2022.12.23.521610

García-Peñalvo, F. J. (2023). La percepción de la Inteligencia Artificial en contextos educativos tras el lanzamiento de ChatGPT: disrupción o pánico. *Education in the Knowledge Society (EKS)*, *24*, e31279. https://doi.org/10.14201/EKS.31279

Gómez, C., y Alende, S. (2022). Rutinas y tareas del profesorado universitario durante la covid-19, Estudio de caso de la facultad de ciencias sociales y comunicación de la Universidad de Vigo. *Investigación: Cultura, Ciencia y Tecnología*, *27*, 36–48. https://dialnet.unirioja.es/servlet/articulo?codigo=8626674&info=resumen&idioma=SPA

Hall, R., Atkins, L., y Fraser, J. (2014). Defining a self-evaluation digital literacy framework for secondary educators: the DigiLit Leicester project. *Research in Learning Technology*, *22*. https://doi.org/10.3402/RLT.V22.21440

Haque, M., Dharmadasa, I., Tasnim Sworna, Z., Namal Rajapakse, R., y Ahmad, H. (2022). "I think this is the most disruptive technology" Exploring Sentiments of ChatGPT Early Adopters using Twitter Data. *Arxiv*. https://doi.org/10.48550/arXiv.2212.05856

Heller, B., Heller, B., Proctor, M., Mah, D., Jewell, L., y Cheung, B. (2005). Freudbot: An Investigation of Chatbot Technology in Distance Education. *EdMedia + Innovate Learning*, *2005*(1), 3913–3918. https://www.learntechlib.org/primary/p/20691/

Khalil, M., y Er, E. (2023). Will ChatGPT get you caught? Rethinking of Plagiarism Detection. *Arxiv*. https://doi.org/10.48550/arXiv.2302.04335

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., y Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, *2*(2), e0000198. https://doi.org/10.1371/JOURNAL.PDIG.0000198

Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales. *Boletín Oficial del Estado, 294,* de 6 de diciembre de 2018. https://www.boe.es/buscar/act.php?id=BOE-A-2018-16673

Locke, S., Bashall, A., Al-Adely, S., Moore, J., Wilson, A., y Kitchen, G. B. (2021). Natural language processing in medicine: A review. *Trends in Anaesthesia and Critical Care*, *38*, 4–9. https://doi.org/10.1016/J.TACC.2021.02.007

Luo, B., Lau, R. Y. K., Li, C., y Si, Y. W. (2022). A critical review of state-of-the-art chatbot designs and applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *12*(1), 1–26. https://doi.org/10.1002/WIDM.1434

Manning, C., y Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press. http://nlp.stanford.edu/fsnlp/

Pérez, J. Q., Daradoumis, T., y Puig, J. M. M. (2020). Rediscovering the use of chatbots in education: A systematic literature review. *Computer Applications in Engineering Education*, *28*(6), 1549–1565. https://doi.org/10.1002/CAE.22326

Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril, relativo a la protección de las personas físicas en lo que respecta al tratamiento de sus datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos). *Diario Oficial de la Unión Europea, L119,* de 4 de mayo de 2016. https://eur-lex.europa.eu/legal-content/es/ALL/?uri=CELEX%3A32016R0679

Rudolph, J., Tan, S., y Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, *6*(1). https://doi.org/10.37074/JALT.2023.6.1.9

Sáiz-Manzanares, M.C., Casanova, J.R., Lencastre, J.A., Almeida, L., y Martín-Antón, L.J. (2022). Satisfacción de los estudiantes con la docencia online en tiempos de COVID-19. *Comunicar*, *30*(70). https://doi.org/10.3916/C70-2022-03

Sancho-Gil, J. M., Sánchez-Valero, J. A., y Domingo-Coscollola, M. (2017). Research-based insights on initial teacher education in Spain. *European Journal of Teacher Education*, *40*(3), 310–325. https://doi.org/10.1080/02619768.2017.1320388

Sharp, J., y Theiler, S. (2018). A Review of Psychological Distress Among University Students: Pervasiveness, Implications and Potential Points of Intervention. *Internatio-*

nal Journal for the Advancement of Counselling, *40*(3), 193–212. https://doi.org/10.1007/s10447-018-9321-7

Tallyn, E., Fried, H., Gianni, R., Isard, A., y Speed, C. (2018). The ethnobot: Gathering ethnographies in the age of IoT. *Conference on Human Factors in Computing Systems - Proceedings*, 604, 1-13. https://doi.org/10.1145/3173574.3174178

Villegas-Ch, W., Arias-Navarrete, A., y Palacios-Pacheco, X. (2020). Proposal of an Architecture for the Integration of a Chatbot with Artificial Intelligence in a Smart Campus for the Improvement of Learning. *Sustainability*, *12*(4), 1500. https://doi.org/10.3390/SU12041500

Zhu, Q., y Luo, J. (2022). Generative Pre-Trained Transformer for Design Concept Generation: An Exploration. *Proceedings of the Design Society*, *2*, 1825-1834. https://doi.org/10.1017/pds.2022.185

Apéndices

Siguiendo las políticas de acceso abierto, se han publicado[7]:

1. Los tres textos elaborados por ChatGPT y los tres textos realizados por alumnos.
2. El instrumento íntegro utilizado en el estudio.

Fecha de recepción: 12 de abril de 2023.
Fecha de revisión: 15 de mayo de 2023.
Fecha de aceptación: 19 de diciembre de 2023.

7 Disponible en: https://osf.io/3xm98/?view_only=426e95fda4394a80b606874e91e25c82