



Education in the Knowledge Society

journal homepage <http://revistas.usal.es/index.php/eks/>

Ediciones Universidad
Salamanca



Can AI fool us? University Students' Lack of Ability to Detect ChatGPT

¿Puede engañarnos una IA? Carencias del estudiantado universitario para detectar ChatGPT

José L. González-Geraldo^{a(*)}, Leticia Ortega-López^b

^a Departamento de Pedagogía, Facultad de Ciencias de la Educación y Humanidades, Universidad de Castilla-La Mancha, Cuenca, España

<https://orcid.org/0000-0003-1698-0122> joseluis.ggeraldo@uclm.es

^b Facultad de Ciencias de la Educación y Humanidades, Universidad de Castilla-La Mancha, Cuenca, España

<https://orcid.org/0009-0001-2041-234X> leticia.ortega1@alu.uclm.es

ARTICLE INFO

Keywords

artificial intelligence; AI; ChatGPT; code interpreter; advanced data analysis; higher education.

Palabras claves

inteligencia artificial; IA; ChatGPT; code interpreter; advanced data analysis; educación superior.

ABSTRACT

The evolution that artificial intelligence (AI) has undergone in recent months, especially in its capacity to generate high-quality argumentative texts, has been a disruptive event in academic environments and higher education spaces. One of the current and future significant challenges we face is the difficulty of identifying those texts that simulate our human narrative in a natural language yet have been crafted by an AI. OBJECTIVES: In the present research, we analyze to what extent university students from degrees in Primary Education and Social Education (n=130) can make this distinction. METHODOLOGY: By implementing ad hoc questionnaires, we verify the degree of perception, complexity, and authorship regarding different texts. The texts to be analyzed were various definitions of the concept of education, half made by humans and the other half by an AI that emulated the degree of complexity and expression of the different profiles and human tones. In parallel, the statistical analyses were conducted using the "Advanced Data Analysis" function (formerly "Code Interpreter") of ChatGPT itself and replicated in SPSS, finding a high similarity between the two, qualitatively consistent in all but one. Additionally, the graphics included were also created using this function. RESULTS: The results indicate the difficulty the students in the sample had in detecting the definitions made by the AI. CONCLUSIONS: Although, as of today, the limits of AI concerning human thought and reasoning are clear, the versatile creative capacity of these language models makes their identification difficult and masks it.

RESUMEN

La evolución que ha sufrido en los últimos meses la inteligencia artificial (IA), especialmente en la capacidad de generar textos de alta calidad argumentativa ha supuesto un hecho disruptivo en ámbitos académicos y en espacios de educación superior. Uno de los mayores retos actuales y futuros a los que nos enfrentamos reside en la dificultad de identificar aquellos textos que simulan nuestra narrativa humana en un lenguaje natural y que sin embargo han sido elaborados por una IA. OBJETIVOS: En la presente investigación analizamos hasta qué punto el estudiantado universitario de los grados en Educación Primaria y en Educación Social (n=130) es capaz de realizar esta distinción. METODOLOGÍA: A través de la implementación de cuestionarios realizados *ad hoc*, comprobamos el grado de percepción, complejidad y autoría con respecto a distintos textos. Los textos a analizar fueron diversas definiciones del concepto de educación, la mitad realizadas por humanos y la mitad realizadas por una IA que emulaba el grado de complejidad y expresión de los distintos perfiles y tonos humanos. De manera paralela, los análisis estadísticos han sido realizados a través de la función "Advanced Data Analysis" (anteriormente "Code Interpreter") del propio ChatGPT y replicados en SPSS, encontrando una gran similitud entre ambos, cualitativamente coincidentes en todos los realizados, excepto uno. Además, los gráficos que se incluyen también han sido realizados gracias a esta función. RESULTADOS: Los resultados señalan la dificultad que los estudiantes de la muestra han tenido en detectar las definiciones realizadas por la IA. CONCLUSIONES: Pese a que, a día de hoy, los límites de la IA con respecto al pensamiento y razonamiento humano son claros, la versátil capacidad creativa de estos modelos de lenguaje dificulta y enmascara su identificación.

(*) Autor de correspondencia / Corresponding author

1. Introducción

Después del proceso de Bolonia, cuya implementación supuso una armonización sin precedentes en la estructura de nuestros sistemas universitarios, muchos nos preguntábamos cuál sería el siguiente gran cambio de paradigma en educación superior. La irrupción de los modelos de lenguaje generativos de inteligencia artificial (IA), como ChatGPT, entre otros, podrían ser la respuesta. Uno de los mayores retos actuales y futuros a los que nos enfrentamos dentro de este tsunami tecnológico, es a la dificultad que tenemos los humanos para identificar aquellos textos argumentativos creados por la IA en un lenguaje natural que simula nuestra narrativa humana, primer objetivo a analizar.

Paralelamente, el uso y las posibilidades de la IA no se reduce a la mera síntesis o generación de ideas. Como quedará corroborado, puede existir una sinergia de estas plataformas con el análisis pormenorizado y detallado desde el punto de vista estadístico, a través de la opción de análisis avanzado (“Advanced Data Analysis” (ADA), anteriormente conocida como “Code Interpreter”), persiguiendo así el segundo de los objetivos. Ambos, por supuesto, íntimamente relacionados y siempre bajo los límites y retos éticos (Carrera Farran & Pérez Garcías, 2023; Future of life, 2023), de selección del contenido sobre el que se basa (Carrasco et al., 2023) o de flagrante menoscabo de la integridad académica (Crawford et al., 2023).

Sin entrar en detalles técnicos, podemos afirmar que ChatGPT es un modelo de lenguaje basado en una arquitectura generativa que utiliza técnicas de aprendizaje profundo y redes neuronales para procesar y generar secuencias de texto de forma autónoma a partir de las entradas del usuario denominadas prompts (Gómez-Cano et al., 2023) que producen respuestas coherentes (Cortés-Osorio, 2023). Esta capacidad es posible gracias al uso de grandes conjuntos de datos de entrenamiento y algoritmos avanzados, que le permiten aprender patrones y relaciones complejas en el lenguaje natural. Además de este entrenamiento no supervisado, existe otra etapa de afinamiento selectivo (*fine-tune*) que convierte a ChatGPT en una herramienta mucho más dialógica que su mero motor GPT.

No hay que olvidar que, a pesar de su enorme potencial para clasificar y relacionar datos, ChatGPT carece de la capacidad de comprender realmente el contexto y la intención subyacente detrás de cada pregunta. Hecho por el cual se ha constatado cómo “alucina” o “delira” (Gravel et al., 2023) imposibilitando de esta forma garantizar resultados de calidad (Alkaissi, 2023). Esta imprecisión, junto con la problemática del todavía escaso límite de la ventana contextual sobre la que puede basarse, normalmente unas 3.000 palabras (Jover et al., 2023), resultan ser el talón de Aquiles de la IA en estos momentos, suponiendo una verdadera amenaza en cuanto a desinformación, desfase de contenido y plagio, entre otros peligros.

Sin embargo, la razón de que este conversador sea capaz de imitar el lenguaje humano con tanta precisión se debe a que la arquitectura de ChatGPT se basa en una red neuronal de transformadores (Srivastava, 2023; Vaswani et al., 2017) que procesa la entrada de texto y aprende a predecir la siguiente palabra o frase (*token*) en función del contexto de la conversación. Este proceso de predicción se basa en la relación entre las palabras y las frases anteriores, lo que unido a una técnica de creatividad aleatoria controlada (*sampling*) le permite generar respuestas relevantes, produciendo así respuestas más diversas y creativas evitando al mismo tiempo la repetición de contenido (Bulatov et al., 2024).

Valorando estas ventanas de oportunidad en espacios de educación superior, La UNESCO ha publicado recientemente una Guía de Inicio Rápido de ChatGPT (Sabzalieva & Valentini, 2023). Esta guía abarca indicaciones sobre cómo darte de alta en la plataforma, qué es y cómo funciona, consideraciones importantes de utilización, e incluso posibles dificultades, como la de evaluar su propia fiabilidad, aspecto en el que se enmarca esta investigación, enlazada con otras ya expuestas en otros foros (González-Geraldo & Ortega-López, 2023; González-Geraldo et al., 2023).

En esta investigación nos centramos en comprobar hasta qué punto podrían pasar inadvertidas creaciones realizadas por una IA (ChatGPT) entre otras elaboraciones humanas similares, todas ellas sobre una misma temática, en nuestro caso, la educativa. De este modo, el principal objetivo de esta investigación radica en el estudio de si, hoy por hoy, la ordenación probabilística de tokens es identificada correcta o incorrectamente por el estudiantado, para, además, observar si existen diferencias significativas con respecto a las variables sexo, grado y la percepción de la muestra en cuanto a grado de acuerdo y complejidad.

De manera paralela a la recopilación de los datos, OpenAI abrió la herramienta denominada “Code Interpreter”, actualmente ADA, basada en el modelo más avanzado hasta la fecha (GPT-4). Esta herramienta permite que el modelo acceda a nuestra propia base de datos y realice complejos análisis con tan solo pedirselo en un lenguaje natural, conversando. Dado que los análisis iban a ser realizados desde un principio a través programas más asentados, decidimos que estos también fueran obtenidos en primer lugar a través de ADA para explorar su fiabilidad y potencial.

Con todo, para sintetizar, los objetivos perseguidos son:

- 1) Estudiar si el estudiantado universitario de los grados de Educación Social y Primaria es capaz de identificar correctamente textos argumentativos complejos generados por ChatGPT frente a textos similares creados por humanos, y observar en qué medida perciben ambos tipos de textos como adecuados — grado de acuerdo— y elaborados —grado de complejidad— en relación con las variables sexo y grado.
- 2) Explorar la viabilidad de utilizar la IA (a través de ADA) para realizar análisis estadísticos específicos de cierta complejidad, en especial en cuanto a contraste de medias no paramétricas, así como otras formas de exposición descriptiva e incluso gráfica.

Con este escenario esperamos abrir ante nosotros una serie de interrogantes tan llenos de inquietudes como de oportunidades. Interrogantes que serán tenidos en cuenta en las conclusiones y que confiamos nos animen a replantearnos no solo lo que nos hace en esencia humanos, sino cuál será el papel que la IA puede desempeñar en la educación universitaria (Romero-Rodríguez et al., 2023) en sus dos facetas principales: la docencia y la investigación.

2. Metodología

Esta investigación responde a las premisas de un estudio cuantitativo de diseño no experimental y de corte transversal a través del uso de cuestionarios tipo Likert (Hernández-Sampieri & Mendoza, 2018).

Para la realización de los análisis se ha llevado a cabo una doble comprobación que nos ha permitido poner a prueba la nueva función ADA de ChatGPT (GPT-4). Durante dicha comprobación se tuvo especial cuidado en no proporcionar en ningún momento datos sensibles que pudieran utilizarse como material de entrenamiento para nuevos modelos. Una vez codificados los datos de manera anónima y estrictamente cuantitativa, repetimos los mismos análisis a través del programa IBM SPSS (v.28). Por motivos de claridad, solo se expondrán y diferenciarán los resultados obtenidos por SPSS en el caso de que estos difieran de los obtenidos por ADA. En caso de no hacerse, se entenderá que SPSS verificó esos mismos resultados. La codificación inicial de los datos fue realizada directamente en SPSS (.sav) mientras que ADA requirió la exportación de estos a un formato compatible (.csv). Todos los gráficos han sido realizados con ADA.

Al no cumplirse el principio de normalidad muestral (Komo-gorov-Smirnov, $p < .001$), se procedió a realizar pruebas de contraste de medias no paramétricas (Mann-Whitney) para así conocer la existencia, o no, de diferencias significativas.

2.1. Muestra

La muestra está formada por alumnado universitario ($n = 130$) de los Grados en Educación Primaria ($n = 43$, 33.1%) y del Grado en Educación Social ($n = 87$, 66.9%) de la Universidad de Castilla-La Mancha (UCLM). El 77.7% ($n = 101$) eran mujeres y el 18.5% ($n = 24$) hombres. Un pequeño porcentaje prefirió no contestar a este ítem (3.8%, $n = 5$). Las personas participantes, mayores de edad, actuaron de forma voluntaria y dieron su consentimiento verbal al comienzo de las mismas, siendo informadas de que los datos recogidos serían tratados de manera anónima y confidencial y que los resultados obtenidos serían publicados y comentados con posterioridad durante las clases participantes, siendo así devueltos como transferencia directa en el impacto de las sesiones docentes y no solo como meros datos de investigación.

El muestreo fue intencionado por conveniencia, ya que buscábamos estudiantes familiarizados con el concepto educativo en todas sus dimensiones: desde el enfoque académico más formal, aportado por los aspirantes a profesores, hacia una concepción más holística dada desde la perspectiva de quienes aspiran a ser educadores y educadoras sociales.

2.2. Instrumento

La recogida de datos se llevó a cabo mediante tres cuestionarios diferentes, que debían ser respondidos por el alumnado de forma independiente y en un orden predeterminado: Cuestionario 1, Cuestionario 2 y Cuestionario 3.

Estos cuestionarios contenían 16 definiciones de educación (Apéndice A). La mitad de estas definiciones fueron creadas por humanos, y la otra mitad por ChatGPT-3.5. Estas mostraban diferentes perfiles y tonos, variando entre elaboraciones complejas, simples, incompletas, poéticas, infantiles, asépticas, profesionales, formales y metafóricas. El punto clave era que la IA imitara los mismos perfiles y tonos humanos, para ello se usaron prompts que seguía el patrón: “En pocas palabras, actúa como X y dame una definición de educación”.

Las definiciones humanas fueron solicitadas expresamente a perfiles previamente seleccionados. Además de estos cinco perfiles, también se añadieron tres pares de definiciones en las que la forma, y no tanto la fuente, fuera emulada: emotiva, no emotiva y metáfora. Cabe señalar que la definición elegida como emotiva es la única no realizada ad hoc para la presente investigación, pues ha de ser atribuida al filósofo Emilio Lledó (ítem 12).

En el Cuestionario 1, el alumnado tenía que mostrar su nivel de acuerdo con las 16 definiciones mediante una escala de Likert con puntuaciones de 1 a 5 (siendo 1 totalmente en desacuerdo y 5 totalmente de acuerdo). La distribución de las definiciones a lo largo de este y el resto de los cuestionarios fue realizada de manera aleatoria, pero en el mismo orden.

En el Cuestionario 2 las personas participantes tenían que indicar la calificación de complejidad que asignaban a las definiciones. Del mismo modo, lo hicieron a través de una escala Likert con puntuaciones de 1 a 5 (siendo 1 muy simple y 5 muy elaborado).

En el Cuestionario 3 el alumnado sólo tenía que marcar qué definición o definiciones creían que habían sido desarrolladas por Inteligencia Artificial y no por un humano. Se advirtió al alumnado de que “alguna” había sido creada por ChatGPT, pero no todas, por lo que disponían de un rango en el que podían marcar un mínimo de una definición y un máximo de 15 definiciones.

Es muy importante destacar el orden en que se administraron los cuestionarios pues hasta la tercera parte no se hizo ninguna mención a la Inteligencia Artificial. Por lo tanto, nos aseguramos de que ambas percepciones -el nivel de acuerdo (Cuestionario 1) y el nivel de complejidad (Cuestionario 2)- se aplicaran sin sesgo posible hacia la fuente de estas definiciones.

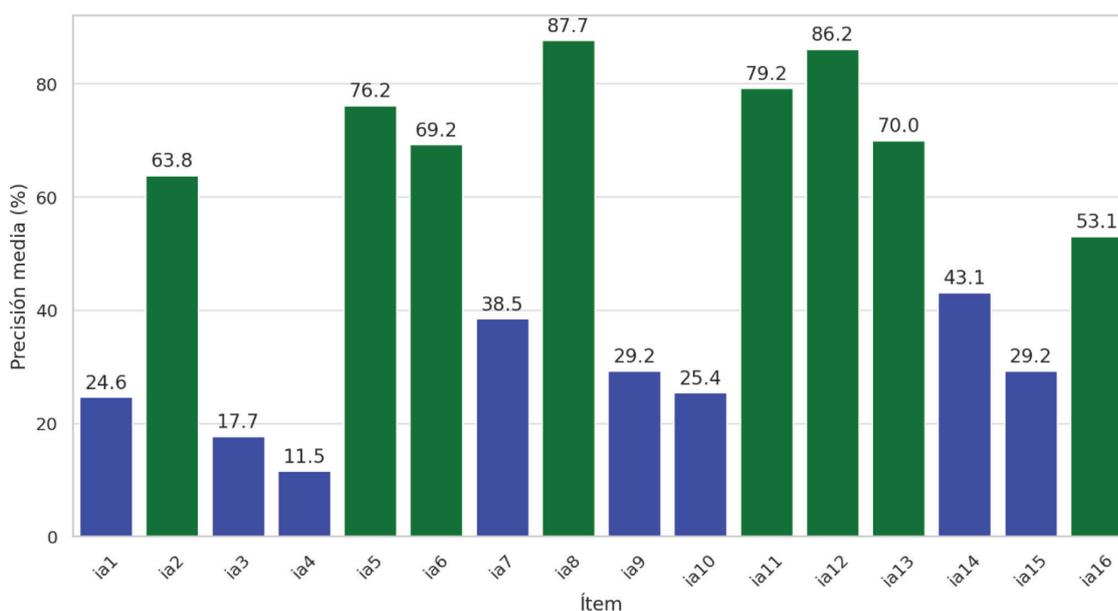
3. Análisis y resultados

3.1. Percepción de autoría en las definiciones

Los resultados muestran que de entre las ocho posibles definiciones de IA presentes, la muestra identificó una media de cuatro definiciones ($\bar{x} = 4.34$), y que de estas cuatro solo dos fueron aciertos reales ($\bar{x} = 2.19$).

Las personas participantes en la muestra ($n = 130$) nunca identificaron ninguna de las definiciones de IA por encima del 50% de acuerdo (Figura 1), presentando una precisión media de acierto del 50.3%, cercana a lo esperado por simple azar (50%).

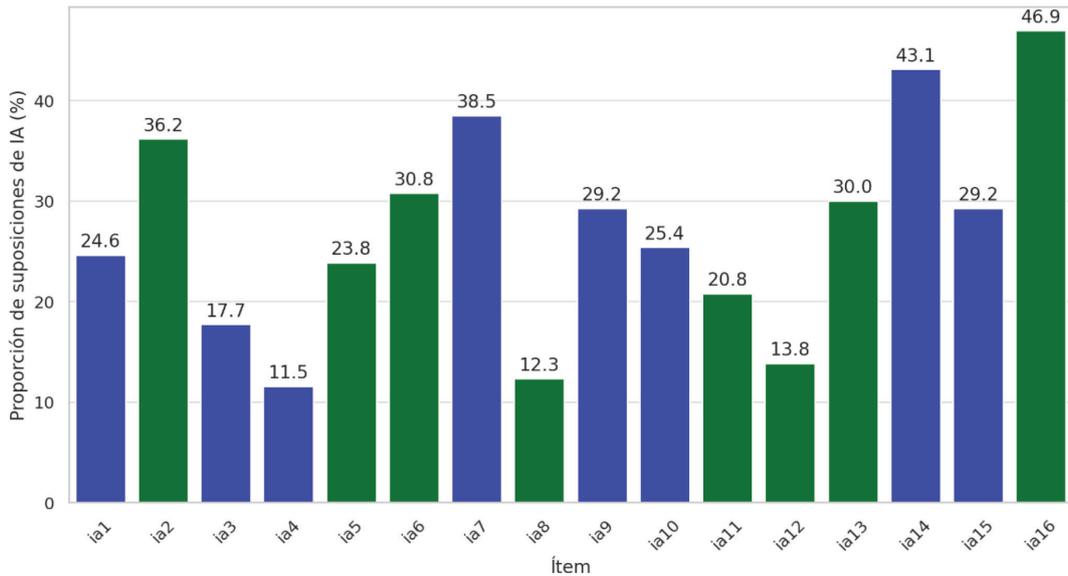
Figura 1. Porcentaje de precisión media por ítem ($n = 130$)



Los datos muestran que el estudiantado otorgó puntuaciones más altas a la calificación de acuerdo con la IA (Cuestionario 1, \bar{x} (IA) = 3.57; \bar{x} (HUM) = 3.09), así como al grado de complejidad de las definiciones de IA (Cuestionario 2, \bar{x} (IA) = 3.17; \bar{x} (HUM) = 2.73). SPSS, por su parte, matizó la media de acuerdo humano (\bar{x} (HUM) = 3.13), así como la del grado de complejidad humana (\bar{x} (HUM) = 2.74).

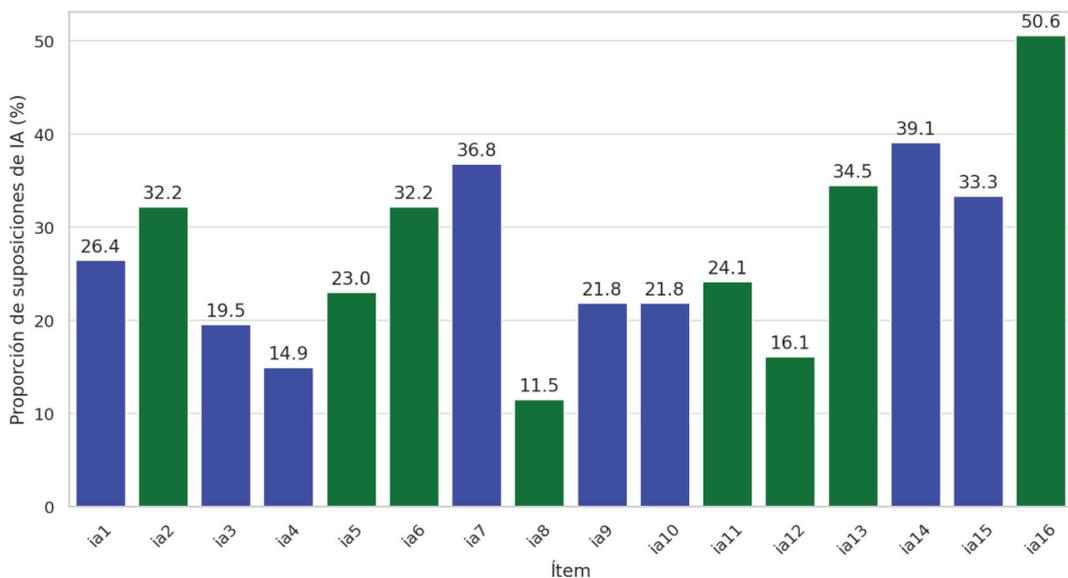
Si nos centramos en comprobar si el estudiantado era capaz de identificar qué definiciones habían sido realizadas por una Inteligencia Artificial, los resultados nos indican que la definición que más cerca estuvo de ser etiquetada por más de la mitad de la muestra como elaborada por una IA fue el ítem 16 (46.9%), seguida de cerca por el ítem 14 (43.1%). Estos dos ítems corresponden, respectivamente, a la definición humana realizada por el profesor universitario de mayor rango de la muestra (Catedrático) y a la respuesta proporcionada por ChatGPT al simular, precisamente, este mismo perfil (Figura 2).

Figura 2. Porcentaje de suposiciones de elaboración de IA por ítem (n = 130)



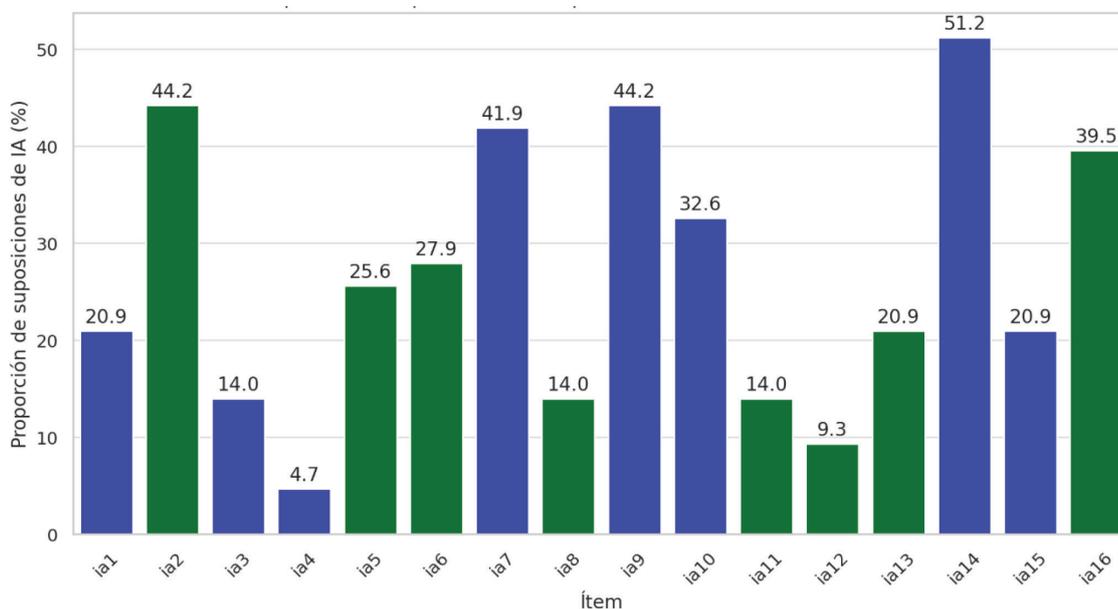
Dividiendo la muestra por grados, observamos algunas apreciaciones relevantes. Atendiendo al Grado en Educación Social (n = 87), estos creen detectar una definición de IA (Figura 3). Esta definición, como ya hemos apuntado fue la elaborada por un profesor Catedrático siendo el ítem 16 (Apéndice A).

Figura 3. Porcentaje de suposiciones de IA. Grado en Educación Social (n = 87)



Atendiendo solo a los futuros maestros y maestras ($n = 43$), los resultados revelan como estos sí que identifican correctamente una de las definiciones realizadas por la IA (Figura 4). Se trata del ítem 14, elaborado por la IA al simular una respuesta propia de un Catedrático de universidad (Apéndice A).

Figura 4. Porcentaje de suposiciones de IA. Grado en Primaria ($n = 45$)



En general no existen diferencias significativas entre el estudiantado de Educación Social y el de Primaria a la hora de detectar correctamente cada una de las definiciones hechas por IA o humanos. En este sentido, únicamente se detectan diferencias significativas en el ítem 9 ($U = 1,452.5$, $z = 2.068$, $p < .05$, $r = .18$), corroboradas también a través de SPSS ($U = 1,452.5$, $z = -2.626$, $p < .05$, $r = -.23$). El ítem 9, recordamos, corresponde a la definición realizada por la IA simulando una escritura emotiva (Apéndice A).

Siguiendo con la variable grado, y aglutinando la tasa de acierto de todas las definiciones realizadas por IA (ítems 1, 3, 4, 7, 9, 10, 14 y 15) y por humanos (ítems 2, 5, 6, 8, 11, 12, 13 y 16) para así realizar un único contraste de medias entre los índices de acierto de cada conjunto, observamos que no existen diferencias significativas con respecto al grado de la muestra ni en la identificación de las definiciones realizadas por IA ($U = 1731$, $z = .69$, $p > .05$) ni en las realizadas por humanos ($U = 1609$, $z = 1.294$, $p > .05$).

En este caso, aunque SPSS corrobora la no existencia de diferencias significativas en cuanto a la identificación de las realizadas por IA ($U = 1,731$, $z = -.713$, $p > .05$), sus resultados sí que indican la existencia de diferencias significativas a favor del Grado en Primaria en relación con la identificación de definiciones humanas ($U = 1,479.5$, $z = -2.003$, $p = .045$, $r = -.13$). Las medianas de ambos grados con respecto a esta tasa de acierto en definiciones humanas, tanto en ADA como en SPSS, es idéntica ($Mdn(ES) = 6 / Mdn(Prim) = 6$).

Por último, atendiendo a la variable sexo y la tasa de acierto en cuanto a la percepción de la autoría de las definiciones, observamos que en general tampoco existen diferencias significativas entre hombres y mujeres. No obstante, existen dos ítems en los que sí que se encuentran: 1) Ítem 14 ($U = 1,525.5$, $z = 1.965$, $p < .05$, $r = .17$), a favor de los hombres, y 2) Ítem 16 ($U = 1,494.5$, $z = 1.771$, $p < .05$, $r = .16$), a favor de las mujeres. Estos resultados también son corroborados a través de SPSS: $U = 898.5$, $z = -2.302$, $p < .05$, $r = -.2$ y $U = 929.5$, $z = -2.046$, $p < .05$, $r = -.18$, respectivamente. Recordemos que el ítem 14 hace referencia a la definición realizada por la IA simulando ser un catedrático de universidad y que el ítem 16 es, precisamente, la definición realizada por el catedrático (Apéndice A).

Al igual que hicimos con la variable grado, aglutinando la tasa de acierto en un único coeficiente observamos que, en conjunto, existen diferencias significativas entre hombres y mujeres -a favor de los hombres ($Mdn(Hom) = 3 / Mdn(Muj) = 2$)- a la hora de identificar correctamente las definiciones realizadas por la IA ($U = 1,582$, $z = 2.319$, $p < .05$, $r = .2$), pero no a la hora de identificar las realizadas por humanos ($U = 1,352$, $z = .878$, $p > .05$). Datos cualitativamente similares en este caso a los encontrados a través de SPSS: $U = 842$, $z = -2.395$, $p < .05$, $r = -.21$ y $U = 1,071$, $z = -.916$, $p > .05$, respectivamente.

3.2. Percepción de acuerdo y complejidad

3.2.1. Diferencias en cuanto al sexo

En relación con el grado de acuerdo, complejidad y la variable sexo, las pruebas Mann-Whitney U realizadas por ADA no revelaron diferencias significativas en las puntuaciones medias de acuerdo y complejidad entre hombres y mujeres. El grado de acuerdo en las mujeres ($Mdn(IA) = 3.63 / Mdn(HUM) = 3.13$) no difiere significativamente del de los hombres ($Mdn(IA) = 3.75 / Mdn(HUM) = 3.13$) en relación con las definiciones elaboradas por IA ($U = 1,278, z = .414, p > .05$) ni en relación con las elaboradas por los humanos ($U = 1,194.5, z = -.11, p > .05$). En cuanto al grado de complejidad detectado por las mujeres ($Mdn(IA) = 3.13 / Mdn(HUM) = 2.63$) no difiere significativamente del de los hombres ($Mdn(AI) = 3.25 / Mdn(HUM) = 2.63$), ni en relación con las definiciones elaboradas por la AI ($U = 1,317.5, z = .661, p > .05$) ni en relación con las elaboradas por los humanos ($U = 1,282, z = .439, p > .05$).

Los resultados obtenidos a través de SPSS coinciden sustancialmente con los presentados por ADA, aunque los detalles de los mismos presentan algunas diferencias de datos precisos. Así, observamos cómo el grado de acuerdo en las mujeres ($Mdn(IA) = 3.63 / Mdn(HUM) = 3.13$) no difiere significativamente del de los hombres ($Mdn(IA) = 3.75 / Mdn(HUM) = 3.13$, en relación con las definiciones elaboradas por la IA ($U = 1,133, z = -.497, p > .05$) ni en relación con las elaboradas por los humanos ($U = 1,204.5, z = -.05, p > .05$). Nuevamente, y en consonancia con el análisis realizado por la IA, los resultados indican que el grado de complejidad en las mujeres ($Mdn(AI) = 3.13 / Mdn(HUM) = 2.63$) no difiere significativamente del de los hombres ($Mdn(AI) = 3.25 / Mdn(HUM) = 2.63$), ni en relación con las definiciones elaboradas por la AI ($U = 1,114.5, z = -.613, p > .05$) ni en relación con las elaboradas por los humanos ($U = 1,132.5, z = -.50, p > .05$).

3.2.2. Diferencias en cuanto al grado

Por otro lado, las pruebas de Mann-Whitney U realizadas por ADA revelaron diferencias significativas en las puntuaciones medias de acuerdo entre el estudiantado del Grado en Educación Social y el Grado en Primaria. Así, el grado de acuerdo de los futuros maestros y maestras ($Mdn(IA) = 3.88 / Mdn(HUM) = 3.38$) difiere significativamente del grado de acuerdo de los futuros educadores y educadoras sociales ($Mdn(IA) = 3.5 / Mdn(HUM) = 3.13$), respecto a las definiciones elaboradas por IA ($U = 1,347, z = -2.59, p < .001, r = -.3$) y con respecto a las definiciones elaboradas por humanos ($U = 1,347, z = -2.59, p < .001, r = -.23$). En ambos casos a favor del Grado en Educación Primaria y con tamaño del efecto pequeño.

Los resultados reportados por SPSS también coinciden de manera cualitativa, aunque con diferencias en cuanto a las puntuaciones obtenidas. En este sentido observamos cómo el grado de acuerdo de los futuros maestros y maestras ($Mdn(IA) = 3.88 / Mdn(HUM) = 3.38$) difiere significativamente del grado de acuerdo de los futuros educadores y educadoras sociales ($Mdn(IA) = 3.5 / Mdn(HUM) = 3.13$) respecto a las definiciones elaboradas por IA ($U = 1,194, z = -3.358, p < .001, r = -.29$) y con respecto a las definiciones elaboradas por los humanos ($U = 1,386, z = -2.403, p < .05, r = -.21$). En ambos casos, de nuevo, el tamaño del efecto es pequeño y a favor del Grado en Educación Primaria.

Paralelamente, tanto en el caso de ADA como SPSS, no se encuentran diferencias significativas en relación con la percepción de la complejidad y el grado del estudiantado.

4. Discusión

A la luz de los resultados obtenidos, observamos cómo la muestra fue incapaz de detectar ninguna de las definiciones realizadas por la IA por encima del 50% de acuerdo que se podría esperar por mero azar (Cuestionario 3).

Existen distintas explicaciones, ciertamente complementarias. Una reside en la complejidad y calidad de los textos elaborados por la IA, cuya naturalidad los hace difíciles de identificar, aunque es preciso recordar que estos textos fueron simulados para que el nivel de complejidad fuera diverso. Otra recae en la corta experiencia de la muestra, estudiantes todavía en formación. El desconocimiento que el estudiantado tiene a la hora de detectar lo que nos gusta denominar la “melodía del algoritmo” (González-Geraldo & Ortega-López, 2023), es decir, a la manera en que estos modelos suelen generar sus resultados, junto con el desconocimiento propio de la disciplina que todavía no manejan como expertos, influyen en el bajo resultado de detección obtenido, cuya media es de dos por sujeto.

Adicionalmente, es interesante señalar que el Cuestionario 3 les pedía explícitamente que señalaran qué definiciones creían que habían sido elaboradas por IA, quedando por defecto todas señaladas, en principio, como elaboración humana. Tal y como demuestran ciertas investigaciones, este hecho no es baladí. Ante las peticiones dicotómicas que requieren cierta complejidad, la opción por defecto suele ser la elegida por los usuarios ante la duda (Amir et al., 2005). Esta teoría respalda que el grado de precisión obtenido por la muestra sea tan elevado en las definiciones realizadas por humanos.

Sin embargo, en cuanto a los grados de acuerdo y complejidad de las definiciones (cuestionarios 1 y 2), cumplimentados sin tener información ni conocimiento sobre el verdadero propósito de esta investigación, los resultados nos indican que la muestra estaba más de acuerdo y consideraba más complejas las definiciones realizadas por ChatGPT. Estos resultados sin duda nos hacen reflexionar sobre lo certera y complejas que pueden llegar a ser las elaboraciones realizadas por una IA que, en verdad, está en sus primeros compases. Independientemente de ello, los hallazgos reportados por ADA y SPSS coinciden al indicar que, independientemente del sexo, las personas encuestadas tienden a estar de acuerdo con las definiciones en la misma medida y perciben las definiciones como igualmente complejas.

Al enfocar el análisis en aquellas definiciones que más cerca estuvieron de ser detectadas como elaboradas -correcta o incorrectamente- como IA, observamos que ambas (ítem 14 (IA-catedrático) e ítem 16 (HUM-catedrático) son las que en principio más complejas y perfectas eran de esperar, tanto en la vertiente artificial como en la humana, y siempre desde el campo de la ciencia y no tanto de la filosofía, donde quizá las metáforas y los vericuetos mentales pudieran ser mejor valorados. El contexto, por tanto, no debe ser ignorado.

Al separar por grados, el estudiantado de Educación Social solo identificó una definición como elaborada por IA, la elaborada por el Catedrático. Prestando atención al Grado en Primaria, comprobamos cómo solamente identifican el ítem 14 (IA-catedrático) como elaborado por una IA. De hecho, este es el único acierto, la única identificación correcta, que hemos podido encontrar. Eso sí, solo al dividir la muestra por grados.

Una posible explicación para estos últimos hallazgos nos lleva a replantearnos si el estudiantado del Grado en Primaria quizá esté más familiarizado con el lenguaje que la propia IA identifica como complejo y, quizá, “formal”, quedando así una definición más cercana al campo profesional en el que un maestro se puede encontrar más cómodo que un educador social. Esta interpretación encajaría, al menos en parte, con las diferencias significativas obtenidas en cuanto al grado, justificando que las personas participantes del Grado en Primaria dieran puntuaciones de acuerdo más altas tanto en las elaboradas por ChatGPT como en las elaboradas por humanos.

De cualquier forma, la inexistencia parcial de diferencias significativas entre grados en cuanto a la tasa de acierto, unido al extremadamente bajo grado de acuerdo encontrado en ambos ítems, al hecho de que solo se haya encontrado al desglosar la muestra -no demasiado amplia- y que el tamaño del efecto de las diferencias reportadas siempre haya sido pequeño, nos hacen decantarnos porque en verdad estamos más ante una simple diferencia porcentual que ante un hecho de calado que pudiera extrapolarse a la población.

Algo distinto ocurre al desglosar por sexo, pues en este caso hombres y mujeres sí que presentan diferencias significativas, precisamente, con respecto a estos dos ítems: 14 (IA-Catedrático), a favor de los hombres, y 16 (HUM-Catedrático), a favor de las mujeres). Al aglutinar la tasa de acierto en un solo índice estas diferencias siguen siendo significativas a la hora de identificar las definiciones realizadas por IA, siendo de nuevo a favor de los hombres, pero no a la hora de identificar las humanas. Estos resultados parecen reforzar el estereotipo machista que asocia el formalismo, conocimiento y uso de las tecnologías a los hombres con respecto a las mujeres, aunque no podemos dejar de lado el sesgo que supone tener una muestra relativamente pequeña en la que menos del 20% de la misma son hombres, por lo que no creemos oportuno aseverar estos resultados de manera categórica, quedando abierta como línea de investigación futura.

En relación con el segundo objetivo de este artículo, el uso y comprobación de las capacidades de análisis de la función ADA, observamos cómo pese a que algunos datos concretos varían, algo asumible al aceptar que la complejidad de los análisis requiere acercamientos no siempre exactos por ambas partes (como ocurre con las correcciones de continuidad de la prueba U), las interpretaciones cualitativas han sido consistentes a lo largo de todo el artículo.

Solamente en un caso ha existido discrepancia, a la hora de detectar diferencias significativas en relación con la tasa de acierto de las definiciones de manera aglutinada y en función del grado. Mientras que ADA no encontró diferencias entre Educación Social y Primaria, SPSS sí que detectó diferencias significativas a favor del estudiantado de Primaria, aunque dichas diferencias se encuentran ciertamente en el límite ($p = 0,045$) y tienen un efecto pequeño ($r = -0,13$). No obstante, esta única discrepancia nos permite comprobar cómo ante esta situación nos vemos abocados a decidir si estamos ante un error de Tipo I (creemos a SPSS cuando en verdad la hipótesis nula es cierta, y no hay diferencias significativas) o de Tipo II (creemos a ADA cuando la hipótesis nula es falsa, y en verdad existen diferencias significativas). En este caso, estando en fase Beta, y dada la larga

trayectoria de SPSS, optamos por asumir el riesgo de caer en un error de Tipo I, de ahí que anteriormente, un par de párrafos atrás, hayamos hablado de “inexistencia parcial de diferencias significativas”.

Por último, queda constancia gráfica de que la capacidad de producir elementos visuales en función de los datos codificados es más que prometedora, sobre todo teniendo en cuenta que, para obtenerlos, simplemente hay que pedirlos en un lenguaje coloquial que, por supuesto, requiere de un conocimiento mínimo y necesario que permita no ya obtenerlos -algo que el propio ChatGPT puede lograr guiando al usuario- sino valorar su pertinencia y precisión.

5. Conclusiones

A la vista de los resultados, que nos ayudan a delinear y delimitar las posibilidades y límites de la IA como parte del fluctuante entramado que constituye la relación entre pedagogía y la tecnología digital (Suárez-Guerrero et al., 2024), en la actualidad no estamos preparados -quizá ni siquiera alerta- para distinguir las creaciones realizadas por la IA frente a las elaboraciones realizadas por humanos. A través de este estudio inicial, pionero en su campo, así se constata al menos en lo que respecta al estudiantado universitario y en este momento. Futuras líneas de investigación podrían centrarse en analizar otros colectivos para ver si estos sí serían capaces de identificar las creaciones elaboradas por la IA.

Por otro lado, que las personas participantes no sean capaces de distinguir las elaboraciones de IA no señala necesariamente la perfección de estas definiciones (recordemos que estas emulaban distintos perfiles y tonos) sino que nos habla del peligro que supone conceder fiabilidad a toda la información que llega hasta nuestras manos, pudiendo caer en la desinformación que nos aporta la IA con sus alucinaciones —entre otras fuentes—. Este hecho apuntala todavía más la amenaza que supone la gran cantidad de información falsa en red (*fake news*) a la que debemos enfrentarnos cada día en el contexto actual de la era digital y para la que la educación superior debe preparar al estudiantado.

Es ciertamente llamativo que al desglosar la muestra cada uno de los grados crea identificar solo una de las definiciones con una tasa por encima del 50%, y que dichas definiciones sean, precisamente, las que en teoría son más complejas: la elaborada por el Catedrático participante en la muestra y la que la IA realizó al suplantar ese mismo perfil. Esto nos hace pensar que es posible que aquello que resume perfección se nos antoje ajeno a la capacidad humana, asociando el error a la persona. Entendiendo esto así, estudiantes mínimamente duchos en elaboración de prompts podrían solicitar de manera explícita alguna que otra errata o fallo gramatical menor a la IA con el fin de enmascarar un fraude académico.

Por otro lado, el hecho de la baja distinción entre creaciones humanas y artificiales por parte del alumnado puede ser extrapolado al otro lado del proceso de enseñanza-aprendizaje. Aquí surgen de manera abrupta innumerables retos para un profesorado que más allá de centrarse en detectar el uso indebido de la IA deben abordar y afrontar nuevos paradigmas de evaluación. De esta forma, la integridad académica y la innovación educativa quedan explicitadas a través de un uso adecuado de la IA.

Paralelamente, la función de análisis de datos ofrecida en fase Beta por OpenAI, ADA, muestra un nivel de precisión y acierto más que prometedor, siendo capaz de ir más allá de los meros análisis descriptivos y realizar pruebas de contraste de medias no paramétricas con estimación del tamaño del efecto con tan solo pedirlo en lenguaje natural. El mismo lenguaje que consigues de calidad que se acercan casi con precisión milimétrica a lo que el investigador realmente quiere. A la luz de estos resultados, es significativo comprobar cómo días antes del envío final de este artículo esta herramienta cambiara su nombre a “Advanced Data Analysis”. Independientemente del nombre que reciba, estamos ante una herramienta cuyos resultados empiezan a ser debatidos y usados con sorprendente aceptación en diversos ámbitos académicos (Huang et al., 2024; Tayebi et al., 2024; Wang et al., 2024), aunque no sin lógicas y necesarias críticas (Taloní et al., 2023) que nos invitan a recordar la imprescindible revisión experta que estos modelos de IA requieren.

Es ciertamente contradictorio, pero a la vez estimulante, comprobar que los resultados nos llevan a pensar que debemos profundizar más en las bellezas y vulnerabilidades que nos hacen humanos, y quizá no confiar tanto en una IA que replica (ChatGPT) quedando así más del lado de una humanidad que transforma. Estimulante en cuanto a posibilidades de superación a las que nos enfrentamos, y contradictorio con respecto a los retos que supone aceptar que no a mucho tardar la docencia y la investigación universitaria se verán abocadas a un uso responsable, pero también ineludible, de estas plataformas.

Arrastrados o embelesados, desde aquella sociedad de la información que dio paso a una sociedad del conocimiento, hoy nos vemos propulsados a lomos de la IA como comburente de la esperada sociedad de la sabiduría, aquella en la que una educación crítica, diferenciadora y diferenciante, jugará un papel decisivo.

Contribución de los autores

José L. González-Geraldo: Conceptualización, Curación de datos, Análisis formal, Investigación, Metodología, Supervisión, Validación, Visualización, Redacción-borrador original, Redacción-revisión y edición.

Leticia Ortega-López: Curación de datos, Análisis formal, Investigación, Metodología, Supervisión, Validación, Visualización, Redacción-borrador original, Redacción-revisión y edición.

Referencias

- Alkaiissi, H., & McFarlane, S. I. (2023). Alucinaciones artificiales en ChatGPT: implicaciones en la escritura científica. *Cureus*, *15*(2). <https://d66z.short.gy/uQM9Bb>
- Amir, O., Ariely, D., Cooke, A., Dunning, D., Epley, N., Gneezy, U., Koszegi, B., Lichtenstein, D., Mazar, N., Mullainathan, S., Prelec, D., Shafir, E., & Silva, J. (2005). Psychology, Behavioral Economics, and Public Policy. *Marketing Letters* *16*, 443–454. <https://doi.org/10.1007/s11002-005-5904-2>
- Bulatov, A., Kuratov, Y., Kapushev, Y., & Burtsev, M. S. (2024). Scaling Transformer to 1M tokens and beyond with RMT. *arXiv*, Article arXiv:2304.11062v2. <https://doi.org/10.48550/arXiv.2304.11062>
- Carrasco, J. P., García, E., Sánchez, D. A., Porter, E., De La Puente, L., Navarro, J., & 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>
- Carrera Farran, X., & Pérez Garcias, A. (2023). Tecnologías digitales en educación: poniendo el foco en la ética. *EduTec, Revista Electrónica De Tecnología Educativa*, (83), 1–6. <https://doi.org/10.21556/edutec.2023.83.2829>
- Cortés-Osorio, J. A. (2023). Explorando el potencial de ChatGPT en la escritura científica: ventajas, desafíos y precauciones. *Scientia et Technica*, *28*(01), 3-5. <https://doi.org/10.22517/23447214.25303>
- Crawford, J., Cowling, M., & Allen, K. A. (2023). Leadership is needed for ethical ChatGPT: Character, assessment, and learning using artificial intelligence (AI). *Journal of University Teaching and Learning Practice*, *20*(3). <https://doi.org/10.53761/1.20.3.02>
- Future of Life (2023). *Pause Giant AI Experiments: An Open Letter*. Future of Life Institute.
- Gómez-Cano, C., Sánchez-Castillo, V., & Clavijo, T. A. (2023). Unveiling the Thematic Landscape of Generative Pre-trained Transformer (GPT) Through Bibliometric Analysis, *Metaverse Basic and Applied Research*, *2*, 33. <https://doi.org/10.56294/mr202333>
- González-Geraldo, J. L., Buedo-Martínez, S., & Ortega-López, L. (2023). Los Micromachismos de ChatGPT: Evidencias sexistas a través de UNICODE, implicaciones socioeducativas. *XXXV Congreso Internacional de la SIPS. Pedagogía Social en una sociedad digital e hiperconectada: desafíos y propuestas*. Salamanca, 4-6 de octubre.
- González-Geraldo, J. L., & Ortega-López, L. (2023). Valid but not (too) reliable? discriminating the potential of ChatGPT within higher education. *International Conference on Education and New Developments (END 2023)*. Lisboa, 24-26 de junio.
- Gravel, J., D'Amours-Gravel, M., & Osmanliu, E. (2023, 2023/09/01/). Learning to Fake It: Limited Responses and Fabricated References Provided by ChatGPT for Medical Questions. *Mayo Clinic Proceedings: Digital Health*, *1*(3), 226-234. <https://doi.org/10.1016/j.mcpdig.2023.05.004>
- Hernández-Sampieri, R., & Mendoza, C. (2018). *Metodología de la investigación. Las rutas cuantitativa, cualitativa y mixta*. McGraw-Hill.
- Huang, Y., Wu, R., He, J., & Xiang, Y. (2024). Evaluating ChatGPT-4.0's data analytic proficiency in epidemiological studies: A comparative analysis with SAS, SPSS, and R. *Journal of Global Health*, *14*. <https://doi.org/10.7189/jogh.14.04070>
- Jover, G., Carabantes, D., & González-Geraldo, J. L. (2023). Asomándonos a la ventana contextual de la Inteligencia Artificial: decálogo de ayuda para la identificación del uso de ChatGPT en textos académicos. *Aula Magna 2.0 [Blog]*. <https://cuedespyd.hypotheses.org/13299>
- Romero-Rodríguez, J., Ramírez-Montoya, M., Buenestado-Fernández, M., & Lara-Lara, F. (2023). Use of ChatGPT at University as a Tool for Complex Thinking: Students' Perceived Usefulness. *Journal of New Approaches in Educational Research*, *12*(2), 323-339. <https://doi.org/10.7821/naer.2023.7.1458>
- Sabzalieva, E., & Valentini, A. (2023). *ChatGPT e inteligencia artificial en la educación superior: Guía de inicio rápido* (ED/HE/IESALC/IP/2023/12). UNESCO e Instituto Internacional de la UNESCO para la Educación Superior en América Latina y el Caribe. <https://bit.ly/3oeYm2f>

- Srivastava, M. (2023). A Day in the Life of ChatGPT as a researcher: Sustainable and Efficient Machine Learning-A Review of Sparsity Techniques and Future Research Directions. *OSF Preprints*. <https://doi.org/10.31219/osf.io/e9p3g>
- Suárez-Guerrero, C., Gutiérrez-Esteban, P., & Ayuso-Delpuerto, D. (2024). Pedagogía digital. Revisión sistemática del concepto [Digital Pedagogy. Systematic Review of the Concept]. *Teoría de la Educación. Revista Interuniversitaria*, 36(2), e31721. <https://doi.org/10.14201/teri.31721>
- Taloni, A., Scordia, V., & Giannaccare, G. (2023). Large Language Model Advanced Data Analysis Abuse to Create a Fake Data Set in Medical Research. *JAMA Ophthalmol*, 141(12), 1174-1175. <https://doi:10.1001/jamaophthalmol.2023.5162>
- Tayebi Arasteh, S., Han, T., Lotfinia, M., Kuhl, C., Kather, J. N., Truhn, D., & Nebelung, S. (2024). Large language models streamline automated machine learning for clinical studies. *Nature Communications*, 15(1), 1603. <https://doi.org/10.1038/s41467-024-45879-8>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gómez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *Conference on Neural Information Processing Systems (NIPS)*. USA.
- Wang, L., Ge, X., Liu, L., & Hu, G. (2024). Code Interpreter for Bioinformatics: Are We There Yet? *Annals of Biomedical Engineering*, 52(4), 754-756. <https://doi.org/10.1007/s10439-023-03324-9>

Apéndice A

Definiciones de educación

Perfil / tono	Humana	IA (ChatGPT)
Niño de 8 años	La educación son palabras que enseñan cosas. (Ítem 13)	La educación es como un juego de construir tu cerebro con nuevos conocimientos y habilidades. Es divertido y te ayuda a crecer y ser más listo. (Ítem 3)
Estudiante (incompleta)	La educación es una herramienta por la cual se transmiten valores, conocimientos, comportamientos, creencias... y que puede ejercerse de manera positiva o de manera negativa, dependiendo del contexto donde te encuentres. (Ítem 5)	La educación es ir a clase, estudiar y obtener buenas calificaciones. Es solo para tener éxito en el futuro. (Ítem 15)
Estudiante (mejorada)	La educación es el ejercicio transformador y liberador que nos posibilita a convertirnos en la versión más plena de nosotros mismos dentro de un entorno e imaginario social concreto. (Ítem 8)	La educación es el proceso de adquirir conocimientos, habilidades y valores a través de la enseñanza y el aprendizaje. Es fundamental para desarrollarse y tener una vida satisfactoria. (Ítem 1)
Profesor	La educación es el proceso a través del cual una persona llega a ser quien puede llegar a ser. (Ítem 6)	La educación es un proceso integral y continuo que involucra el desarrollo cognitivo, emocional y social. Busca fomentar la curiosidad, el pensamiento crítico y el respeto hacia los demás y el mundo. La educación también prepara a las personas para ser ciudadanos activos y responsables. (Ítem 7)
Catedrático	Proceso intelectual continuo y voluntario por el que una persona progresa gracias a los saberes previamente alcanzados por otros y gracias a los cuales pueden crear nuevo conocimiento. (Ítem 16)	La educación es un proceso de adquisición de conocimientos, habilidades y valores a través de experiencias formales e informales. La educación superior es un nivel más avanzado de educación, con un enfoque en la investigación y el desarrollo de habilidades críticas y de pensamiento. (Ítem 14)
Emotiva	La educación no es una forma de ganarse la vida, es una forma de ganar la vida de los demás. (Ítem 12)	La educación es un proceso de adquisición de conocimientos, valores y habilidades que transforma a las personas y les ayuda a desarrollarse como individuos y a contribuir a la sociedad. Conlleva emoción, motivación y desafío, y puede ser gratificante, inspirador y transformador. (Ítem 9)
No emotiva	La educación es un proceso de adquirir conocimientos, habilidades y valores a través de la enseñanza y el aprendizaje. Es fundamental para desarrollarse y tener una vida satisfactoria. (Ítem 2)	La educación es un proceso de adquisición y desarrollo de conocimientos, habilidades y valores. (Ítem 10)
Metáfora	La educación es fuente que alimenta, brazo que acompaña, espina que incomoda, faro que ilumina, fuego que calienta, llave que abre mundos y alas para soñar. (Ítem 11)	La educación es como un jardín que se cultiva con paciencia y cuidado, creciendo flores de conocimiento y sabiduría. (Ítem 4).