IES PLAZA DE LA CRUZ

upna Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

CHATGPT EN EDUCACIÓN.

UN ANÁLISIS CRÍTICO DE LA INTELIGENCIA ARTIFICIAL GENERATIVA

PROYECTO DE:
Yasmin Bouguetoch y Eve Igbinadolor

2024-2025

TUTORA IES:
Iris Sancho
TUTORES UPNA
Laura de Miguel y Iosu Rodríguez

Resumen

El presente proyecto de investigación aborda la aplicación del modelo ChatGPT en diversas áreas, con un énfasis especial en el educativo. La primera parte del estudio analiza las respuestas del equipo docente y el alumnado, obtenidas a través de un formulario, en relación al uso y percepción que tienen de la herramienta. El objetivo es comprender el impacto real de esta tecnología en el contexto educativo. Asimismo, en la segunda parte del proyecto se somete a ChatGPT-3.5 y sus posteriores versiones a determinados exámenes de las asignaturas de física y filosofía. Finalmente, se evalúan y comparan los diferentes modelos siguiendo una rúbrica de evaluación y extrayendo conclusiones sobre el funcionamiento y las capacidades que presentan las distintas versiones de la herramienta.

Palabras clave: ChatGPT, educación, modelos generativos de lenguaje humano

Abstract

This research project addresses the application of the ChatGPT model in various areas, with a special emphasis on education. The first part of the study analyzes the responses of the teaching team and the student, obtained through a form, in relation to the use and perception they have of the tool. The objective is to understand the real impact of this technology in the educational context. Likewise, in the second part of the project, ChatGPT-3.5 and its subsequent versions are subjected to certain exams in the subjects of Physics and Philosophy. Finally, the different models are evaluated and compared following an evaluation rubric and drawing conclusions about the operation and capabilities presented by the different versions of the tool.

Key words: ChatGPT, education, human language generative models

ÍNDICE

1. Justificación

ChatGPT constituye una revolución en todos los ámbitos de la vida, abriendo muchas puertas y a su vez planteando muchas preguntas en una sociedad todavía no adaptada. Sin embargo, la oposición a su avance no parece ser una solución y se observa cómo se plantean retos ante aplicaciones cada vez más comunes de la IA (inteligencia artificial), como los deep fake o la automatización del periodismo y la guerra (Holmes et al., 2021).

Se ve claramente plasmado el interés que esta inteligencia artificial suscita en la sociedad y por tanto la importancia de estudios al respecto. Y es que ChatGPT en tan solo 5 días alcanzó el millón de usuarios. «Según un informe publicado por el portal australiano *Financial Review*, a Netflix le tomó 40 meses alcanzar esa cifra; a Twitter, 24 meses; a Facebook, 10; y a Instagram, tres» (Diego et al., 2023, p. 6).

En el ámbito educativo se encuentra un vacío respecto a esta herramienta. Es fundamental según la UNESCO (Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura):

Considerar la posibilidad de elaborar mecanismos de seguimiento y evaluación para medir el impacto de la inteligencia artificial en la educación, la docencia y el aprendizaje, a fin de proporcionar una base válida y sólida basada en datos empíricos para la formulación de políticas (UNESCO, 2019, p. 37).

También es importante según el punto 18 del Consenso de Beijing:

Establecer planes a medio o largo plazo y adoptar medidas urgentes para apoyar a las instituciones de educación superior y de investigación en la elaboración o la mejora de cursos y programas de investigación para desarrollar el talento local en materia de inteligencia artificial, a fin de contar con un gran número de profesionales locales de la inteligencia artificial que tengan la pericia necesaria para diseñar, programar y elaborar sistemas de inteligencia artificial (UNESCO, 2019, p. 34).

Se puede observar la puesta en marcha de dichas directrices dentro de la educación pública española en casos como el de la Junta de Andalucía donde en el punto 3.2 de su estrategia con respecto a la inteligencia artificial: busca «facilitar la formación de talento especializado en IA en una medida no tan solo a corto plazo» (Estrategia Andaluza de Inteligencia Artificial 2030, s. f.).

En definitiva, la IA puede ser una herramienta poderosa que permite abrir la caja negra del aprendizaje humano al tener la capacidad de proporcionar una comprensión profunda y detallada de cuándo y cómo éste ocurre realmente, influido por el contexto del alumno (Luckin y Holmes, 2016). Por ello, el presente trabajo pretende realizar un análisis del progreso entre distintos modelos de ChatGPT y su desempeño en las materias de física y filosofía. Se trata de brindar una visión clara sobre el funcionamiento de esta herramienta y sus capacidades. Además, este trabajo permitirá conocer también la opinión y el uso de esta herramienta por parte de la comunidad educativa. Asimismo, se pretende recoger información sobre los principales debates éticos y el impacto social y ecológico que tienen las IA a fin de colaborar con la concientización en el tema.

A nivel personal, el uso generalizado, que ya no se limita a instituciones y brinda acceso práctico a cualquier persona con un ordenador o teléfono; la rapidez incesante con la que avanza esta tecnología; y los logros impresionantes que alcanza son objeto de reflexión y motivan este proyecto. Un ejemplo es la IA *Alphafold*, desarrollada por Google *DeepMind*, que predice la estructura tridimensional de una proteína a partir de su secuencia de aminoácidos y ya ha sido Premio Nobel de Química en 2024 (Parra, 2024).

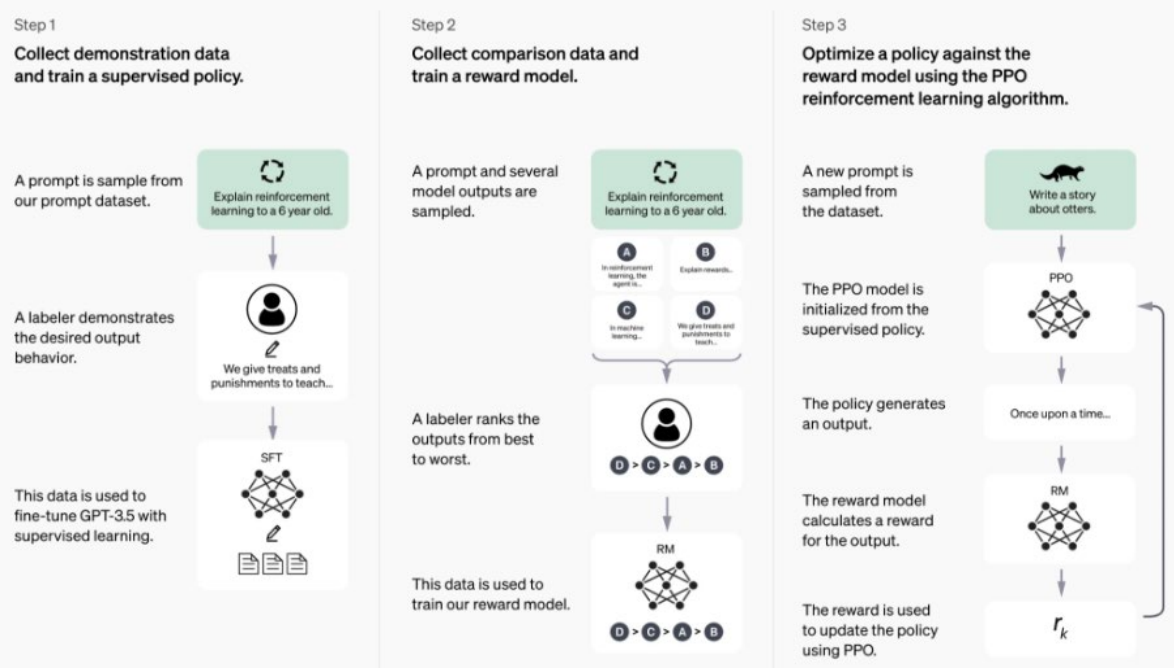
2. Objeto de estudio

El presente trabajo se centra en el modelo generativo de lenguaje ChatGPT. Los modelos *GPT (Generative Pre-trained Transformer o* Transformador Pre-entrenado Generativo) están desarrollados para generar texto natural coherente, ya sea en forma de frases, párrafos o documentos completos de manera consistente con el lenguaje humano. Su característica principal radica en la capacidad de pre-entrenarse con grandes volúmenes de datos textuales y luego ajustarse a tareas específicas adicionales como la clasificación de texto o la respuesta a preguntas. Durante el pre-entrenamiento el modelo aprende a predecir la siguiente palabra en una secuencia de texto sin requerir etiquetas explícitas en los datos de entrenamiento. Esta fase es esencial para que el modelo comprenda y generalice patrones lingüísticos tales como la sintaxis, gramática y semántica. Tras el pre-entrenamiento el modelo puede ser especializado en tareas particulares al proporcionarle un conjunto de datos más reducido y etiquetado.

En la página *web* de OpenAI se detalla el método utilizado para entrenar ChatGPT. Este proceso se basa en el Aprendizaje por Refuerzo con Retroalimentación Humana (*RLHF*, por sus siglas en inglés). En la Figura 1 se muestra el proceso seguido para el entrenamiento del modelo, explicado de manera visual.

Figura 1

Método de entrenamiento de ChatGPT.



Fuente: Adaptado de *OpenAI (2024). https://www.openai.com*

En la figura 1, se puede ver que el entrenamiento consta de tres pasos principales. En el primer paso, se recopila información y se entrena al modelo inicial mediante supervisión humana, donde entrenadores proporcionan ejemplos de conversaciones simuladas. En el segundo paso, las respuestas generadas por el modelo se refinan y se evalúan nuevamente por humanos, clasificándolas de mejor a peor para crear un modelo de recompensa. Finalmente, este modelo de recompensa se utiliza para ajustar el comportamiento del modelo a través de un proceso iterativo de aprendizaje por refuerzo.

Las diferentes versiones de ChatGPT han evolucionado significativamente en términos de capacidad y funcionalidad. Desde su versión inicial en 2018, cada modelo ha aumentado en tamaño y habilidades. Por ejemplo, GPT-3 introdujo capacidades avanzadas de lenguaje, mientras que GPT-3.5 se centró en aplicaciones más amplias, como la programación y el diseño. En marzo de 2023, ChatGPT-4 amplificó enormemente su tamaño y se enfocó en mejorar la seguridad y utilidad de las respuestas generadas (Diego et al., 2023). Además, se inauguró una versión más avanzada de ChatGPT-4 denominado ChatGPT.4o el 13 de mayo de 2024. A diferencia de ChatGPT-4, este último es más eficaz en sus capacidades de texto, voz y visión (OpenAI, 2024). Además, en septiembre de 2024 se introdujo una nueva versión: ChatGPT o1-preview. Esta versión tiene una mejor capacidad de razonamiento que el modelo GPT-4 y por ello es más

lento a la hora de responder al usuario. Asimismo, tiene mejores niveles de razonamiento en física, química y biología (Temsah et al., 2024).

Por otro lado, se hace uso de ChatGPT-4.o en *you.com* para comparar las respuestas del mismo modelo de IA en distintas *webs*. *You.com* es un buscador alternativo de Google con inteligencia artificial generativa. Este buscador dispone de distintos modelos de IA, gratuitos para los usuarios (Codina, 2023).

Para realizar el proyecto, se analizan las respuestas recibidas en función a diferentes *prompts* en modelos de examen de Filosofía o de física por ChatGPT-3.5, ChatGPT-4, ChatGPT-4o y ChatGPT-o1 y se comparan. Los tres primeros mencionados en Filosofía y los 4 en física, además se evalúan las respuestas recibidas en *You.com* también.

A su vez, se realiza una encuesta al alumnado del IES Plaza de la Cruz y al equipo docente sobre el uso de esta IA en el ámbito educativo.

3. Antecedentes

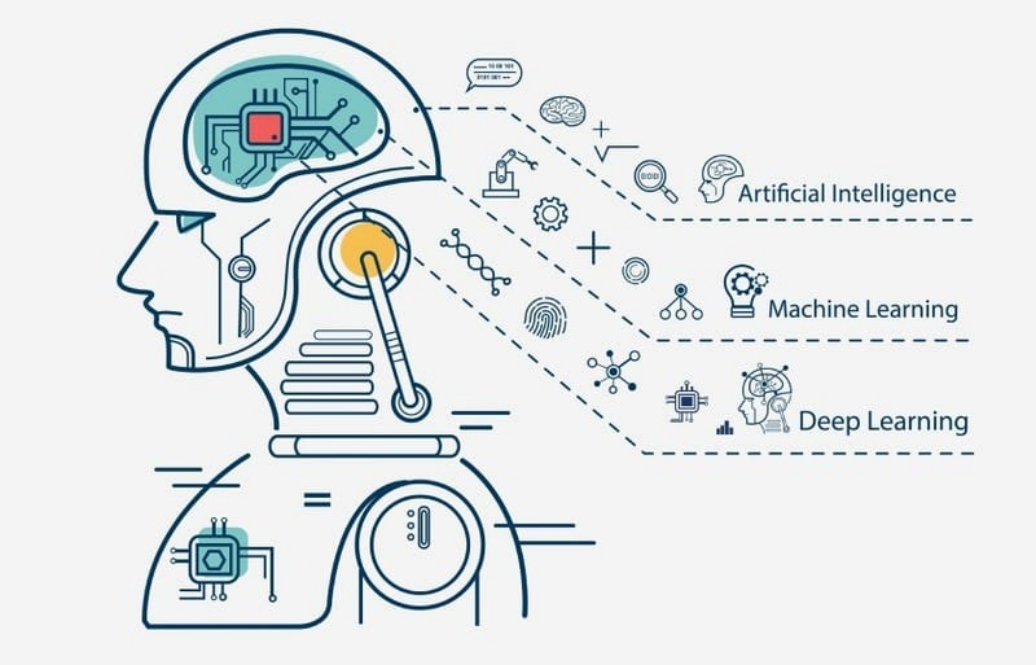
3.1. Definiciones operacionales

Aprendizaje automático o *Machine learning:* es un subconjunto de IA que de forma analítica, capacita a las computadoras para aprender sin ser programadas explícitamente para ello utilizando algoritmos para identificar patrones en grandes conjuntos de datos (Rouhiainen, 2018).

Aprendizaje profundo o *Deep Learning:* es un subcampo del aprendizaje automático que se emplea para abordar problemas altamente complejos que generalmente involucran grandes conjuntos de datos. Este enfoque se basa en el uso de redes neuronales las cuales están estructuradas en capas para identificar relaciones y patrones sofisticados en los datos. Su implementación demanda una amplia cantidad de información y una capacidad de procesamiento computacional considerable (Rouhiainen, 2018).

Figura 2

Machine learning



Fuente*:* Adaptado de Machine Learning in the Real World | Blog, s. f. https://www.insiris.com/blog/machine-learning-in-the-real-world

Computación tradicional/IA Clásica: se caracteriza por la escritura de secuencias de reglas condicionales *'IF'...* (SI) *'THEN'...* (ENTONCES) y otras lógicas condicionales que la computadora sigue para llevar a cabo una tarea.

Datos estructurados: incluyen información concreta como valores numéricos, fechas, monedas, etc… (Rouhiainen, 2018).

Datos no estructurados: incluye información compleja de examinar como imágenes, videos y textos (Rouhiainen, 2018).

IA: Inteligencia artificial

Problemas inversos o *inverse problems:* son situaciones en la que se intenta determinar las causas o condiciones que dan lugar a ciertos efectos observados. Por ejemplo, Newton encontró que el producto entre masa por aceleración, era gravedad. A partir de observar los efectos, dedujo la relación. Lo contrario a los problemas directos o *direct problems*, en los que a partir de la causa se predice el efecto.

Prompt: término informático que se refiere a una instrucción que se le proporciona a un modelo de inteligencia artificial para que genere una respuesta o salida específica. Cuanto más específico es un prompt, más eficaz es. Un *prompt* muy eficaz se compone de contexto (indicación de cómo debe comportarse el modelo), *input* (instrucción previa sobre la información que se va a aportar al modelo), objetivo (instrucción al modelo sobre lo que se quiere que otorgue), limitantes o detalles (otras especificaciones sobre cómo debe ser el resultado) y finalmente la petición (se especifica el *input*). La ingeniería de *Prompts* o *Prompt Engineering* es la técnica utilizada en el aprendizaje automático y el procesamiento del lenguaje natural para

buscar mejorar los resultados que se obtienen mediante la formulación y modificación de las instrucciones de entrada siguiendo los parámetros explicados anteriormente.

Red neurológica artificial (RNA): es un modelo de inteligencia artificial inspirado en la estructura de las redes neuronales biológicas. Compuesta por capas de entrada, capas ocultas para procesamiento intermedio y una capa de salida, las RNA ajustan las conexiones entre neuronas durante el aprendizaje por refuerzo y retropropagación (Holmes et al., 2021).

Token: Los *tokens* representan secuencias de caracteres comunes. 1 *token* equivale a aproximadamente 4 caracteres o 0.75 palabras en texto en inglés.

3.2. Contextualización

ChatGPT es una herramienta de lenguaje basada en inteligencia artificial y desarrollada por OpenAI que utiliza algoritmos de aprendizaje automático para generar textos que imitan el lenguaje humano (Ariyaratne et al., 2023). Debido a su accesibilidad y utilidad, ocasiona ciertos cambios en diversos ámbitos como en el de la educación.

3.2.1. IA en la actualidad; un enfoque en ChatGPT

En noviembre de 2022, se lanzó ChatGPT-3.5. Este modelo fue estudiado al inicio de este trabajo junto a GPT-4. Pero, ¿Qué diferencias existen entre los diferentes *GPT*? En la siguiente tabla a partir de (Diego et al., 2023) se analizan los aspectos más relevantes:

Tabla 1

Comparación entre las diferentes versiones de ChatGPT

Versión	Lanzamiento	Características y/o funciones
GPT	2018	Tenía 117 millones de parámetros.[a]
GPT 2	2019	Tenía 1,5 mil millones de parámetros.[a]
GPT 3	2020	Tenía de 12 a 96 capas y llegó a 175000 millones de parámetros. Podía procesar hasta 2,049 *tokens*.[a]
GPT 3.5	noviembre 2022	Un modelo de lenguaje autorregresivo, abierto, público y gratuito, y capaz de programar, diseñar, «hablar» de política, economía y disímiles temas. Podía procesar hasta 4,096 *tokens* como entrada.[a]
GPT 4	15 de marzo de 2023	Se basa en 100 trillones de parámetros, casi 600 veces más que sus antecesores. Sus capacidades están directamente relacionadas al lenguaje para que logre respuestas más seguras y útiles; mayor precisión en tareas como generación. Podía procesar hasta 8,192 *tokens* como entrada en sus primeras versiones. La versión gpt-4-turbo-2024-04-09 puede procesar

		128000 *tokens* de entrada.[a]
GPT 4o	13 de mayo de 2023	*OpenAI* no ha revelado oficialmente el número de parámetros exactos en los que se basa. Puede procesar hasta 128000 *tokens* de entrada.[a] Este modelo acepta como entrada cualquier combinación de texto, imagen, audio y video, y genera como salida cualquier combinación de estos. Puede responder a entradas de audio en tan solo 232 milisegundos, con un promedio de 320 milisegundos, lo cual es similar al tiempo de respuesta humana en una conversación. [b]

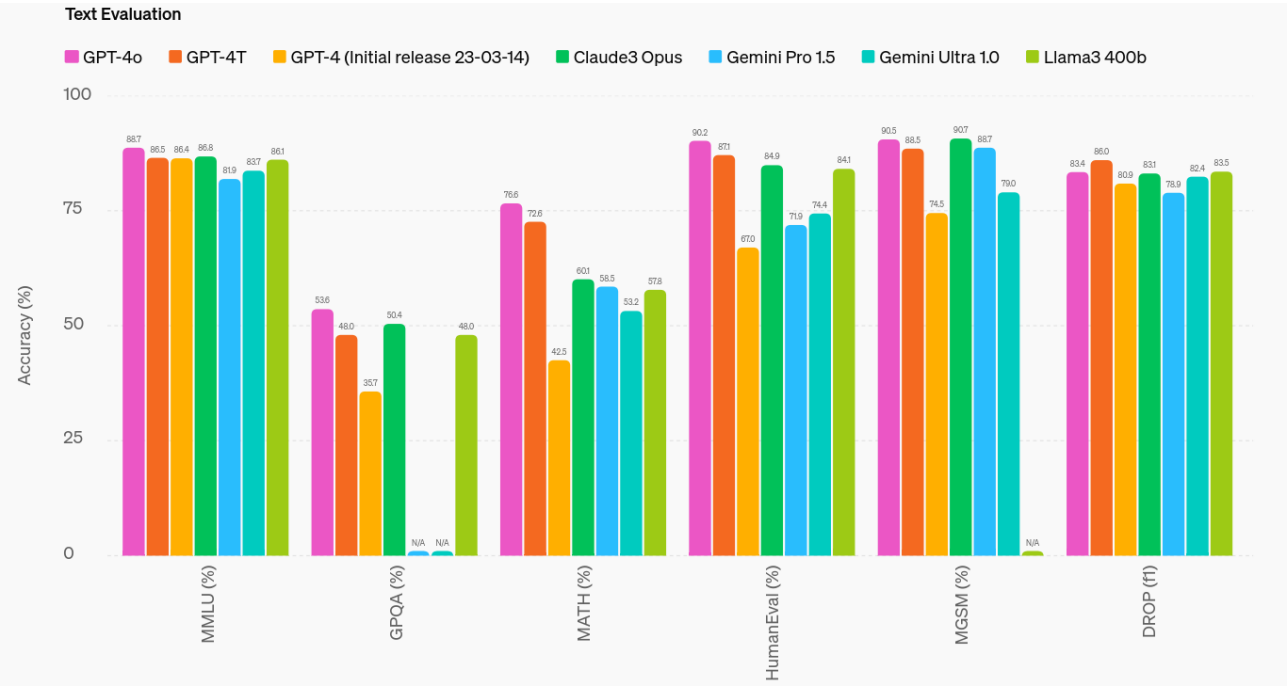
Nota. Adaptado de «ChatGPT: origen, evolución, retos e impactos en la educación», por Diego, F., Morales Suárez, I., & Vidal Ledo, M., 2019[a], Educación Médica Superior, 37(2), p. 8. CC BY-NC y OpenAI (2024)[b].

Cabe destacar que en el término ChatGPT-4o, la «o» representa «omni». Además, antes de GPT-4o, el modo de voz permitía hablar con ChatGPT con latencias promedio de 2.8 segundos (GPT-3.5) y 5.4 segundos (GPT-4). Los anteriores usaban modelos separados para procesar audio y texto, en cambio este último integra todo en un único modelo, lo que le permite reconocer tonos, múltiples hablantes o generar expresiones.

En el siguiente gráfico se presenta una comparación del rendimiento entre diferentes modelos de IA en tareas de comprensión general del lenguaje y razonamiento.

Figura 3

Desempeño de diferentes modelos de IA generativa según 6 'benchmarks'.



Fuente: Adaptado de OpenAI (2024). https://openai.com/index/hello-gpt-4o/

En la figura 3 aparecen distintos *benchmark*. Estos son marcos de prueba estandarizados diseñados para evaluar cómo de bien los modelos de lenguaje se desempeñan en diferentes tareas y capacidades (Patel, 2024). Se puede observar como ChatGPT-4o y ChatGPT-4 Turbo son los que mejor desempeño tienen.

- El *MMLU (Massive Multitask Language Understanding)* o (Comprensión masiva multitarea del Lenguaje) evalúa a los modelos en diferentes materias que van desde Matemáticas básicas hasta derecho profesional (Patel, 2024).
- El *GPQA (Graduate-Level Problem Solving and Question Answering)* o Resolución de Problemas y Respuestas a Preguntas de Nivel de Posgrado, representa el desempeño en preguntas de opción múltiple complejas en áreas como biología, física y química (Patel, 2024).
- El *MATH (Measuring Mathematical Problem Solving with the MATH Dataset)* o Evaluación de la Resolución de Problemas Matemáticos con el Conjunto de Datos MATH consiste en la capacidad de resolver problemas matemáticos (Patel, 2024).
- El *HumanEval* (que proviene de la frase evaluación humana) mide las capacidades de generación de código a partir de descripciones en lenguaje natural humano (Patel, 2024).
- El *MGSM (Multilingual Grade School Math)* o Matemáticas de Escuela Primaria Plurilingüe evalúa la capacidad de los modelos para resolver problemas matemáticos en 10 idiomas diferentes. Utilizando 250 problemas de nivel escolar (Patel, 2024).
- EL *DROP (Discrete Reasoning Over Paragraphs)* o Razonamiento Discreto sobre Párrafos, mide la comprensión lectora y la capacidad de razonamiento discreto de los modelos, requiriendo que extraigan y manipulen información de párrafos para responder preguntas complejas.

3.3. Estado de la cuestión

En el año 2023, se realizó un estudio que analizó los desafíos e impactos potenciales derivados de los Generadores de Texto con Pre-entrenamiento (*GPT*) en diversos campos. Este estudio concluyó que era pertinente abordar el tema centrándose en las oportunidades y desafíos que este plantea para la educación y el aprendizaje, con especial atención a su aplicación en la enseñanza médica (Diego et al., 2023). No obstante, se observan puntos relevantes que no fueron abordados. Por ejemplo, ¿cuál es el margen de error que presenta esta nueva tecnología y cuál es su eficacia en la formación de alumnos, profesores…?

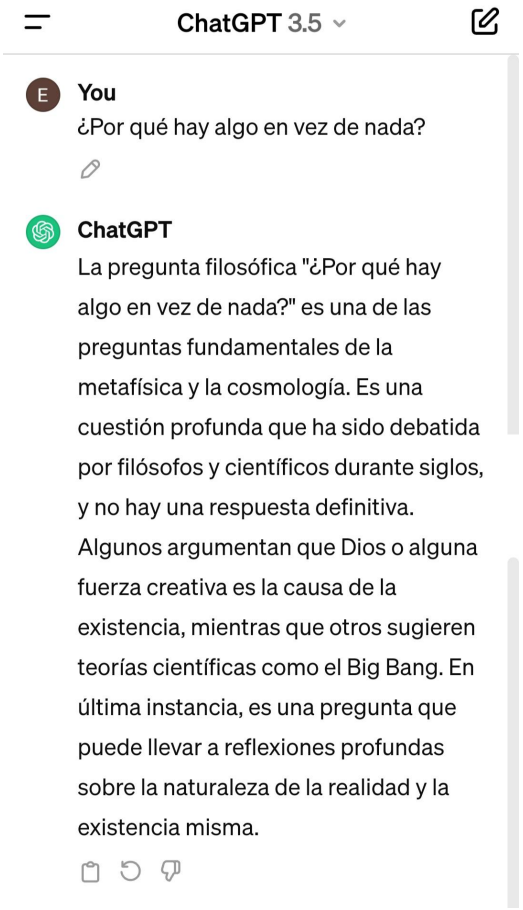
Además, en enero de 2023, Mehmet Firat publicó un artículo titulado «*How chat GPT can transform autodidactic experiences and open education?*» (¿Cómo puede ChatGPT transformar las experiencias autodidactas y la educación abierta?) detallando cómo funciona ChatGPT con datos proporcionados por los creadores. Sin embargo, no se explicó cómo el modelo relaciona los patrones y gestiona los datos para llegar a una respuesta concreta debido al propio desconocimiento concerniente al tema.

El 24 de mayo del 2023 Graziella Orrù, publicó un artículo en el que se comparó los resultados obtenidos por ChatGPT solucionando determinados problemas que requerían de una buena comprensión lectora, con las respuestas proporcionadas por humanos, analizando también el número de respuestas correctas. ChatGPT demostró potencial para resolver problemas que una persona normalmente afrontaría por el método de resolución *insight*. Es decir, en un momento de claridad repentina (Orrù et al., 2023). El número de respuestas correctas que se obtuvo fue casi equivalente al promedio humano.

Considerando que esta tecnología se fundamenta en redes neuronales entrenadas para predecir la salida verbal más probable basándose en un cierto orden de palabras, en este estudio se destacó el debate sobre si estas IA realmente poseen habilidades para resolver problemas, o si tales habilidades son el resultado de una comprensión profunda del problema. El presente trabajo es relevante a la hora de comprender cómo se desenvuelve el modelo en este aspecto, puesto que algunos inconvenientes del trabajo «*Human-like problem-solving abilities in large language models using ChatGPT*» (Capacidades de resolución de problemas similares a las humanas en modelos de lenguaje grandes utilizando ChatGPT) son que solo estudiaba un tipo de problema, empleaba el modelo ChatGPT-3 y la muestra humana era muy reducida. En este contexto, el presente estudio analiza modelos más recientes en el área de la filosofía y la física. También se destaca, la necesidad de realizar un experimento en el que ChatGPT sea capaz de resolver ejercicios orientados a un ámbito un tanto más subjetivo y reflexivo (Figura 4). De esta manera, se podrá evaluar esta IA de una forma más detallada.

Figura 4

Usuario realiza una pregunta subjetiva a ChatGPT.



Fuente: Elaboración propia.

En un estudio realizado por Habib et al., se planteó el impacto de la inteligencia artificial, específicamente ChatGPT 3, en la creatividad de los estudiantes (2024). El estudio abogó por un enfoque cuidadoso en la integración de la IA en la educación creativa, enfatizando la importancia de equilibrar la creatividad humana y la inteligencia artificial. En definitiva, sugirió la necesidad de estrategias pedagógicas enfocadas en la integración cuidadosa de recursos de IA combinados con metodologías de enseñanza tradicionales para fomentar la confianza creativa y las habilidades de pensamiento divergente y convergente en los estudiantes. Esto es especialmente relevante y mencionado en el Anexo D del presente documento.

3.3.1. La cara oculta de la programación. Quién programa, para qué programa, desde dónde programa.

Desde que Donna Haraway escribiera en 1991 «Ciencia *cyborgs* y mujeres» estableciendo el concepto de «conocimiento situado» no se puede hacer un análisis serio sobre el pensamiento artificial sin tener en cuenta de dónde nace.

Tal como describe Haraway:

> El conocimiento situado propone hablar de los objetos de estudio poniendo en evidencia el lugar desde el cual se parte, ya que, independientemente del tipo de método empleado, ningún conocimiento está desligado de su contexto ni de la subjetividad de quién lo emite.

Así, se hace necesario conocer y utilizar esta epistemología en el estudio de la IA, ya que la realidad puede observarse desde muchos lugares. Y es un dato harto relevante mostrar la perspectiva desde la que miramos, de manera que el conocimiento siempre será parcial y situado. Sólo teniendo en cuenta un amplio conjunto de miradas y perspectivas podremos tener un conocimiento más cercano o profundo de la realidad.

En textos posteriores, Haraway también propone el término de objetividad radical (Prasad et al., 2020) en los que se asume la parcialidad del conocimiento, al mismo tiempo que se asume la validez de ese

conocimiento en tanto es real desde el punto de vista de quien lo produce. Esta postura trata de romper la desigualdad que se da entre los sujetos dignos de poseer conocimiento y aquellos que no, denominados como subalternos en las teorías de feminismo postcolonial (Spivak, 2022).

De esta manera toda investigación es una forma de difracción (producir teoría y ahondar en su complejidad) en lugar de la reflexión (representación de la realidad).

Así, es importante, abordar las cuestiones éticas que aparecen con el acceso a la IA. Un artículo muy relevante publicado en 2024 por Aguilar et al., señala que se debe mirar hacia el futuro y asegurarse de que los avances tecnológicos estén respaldados por valores de justicia, inclusión e integridad ética, subrayando que los maestros deberían desempeñar un papel crucial en la toma de decisiones relacionadas con la IA en el ámbito educativo. Sin embargo, su efecto en la educación no es el único factor de preocupación. En un estudio publicado en 2023 por Ray titulado «*ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems*» (ChatGPT: Una revisión exhaustiva sobre antecedentes, aplicaciones, desafíos clave, sesgos, ética, limitaciones y alcance futuro. Internet de las Cosas y Sistemas Ciberfísicos), se mencionan algunos de los retos que esta tecnología plantea: la fiabilidad y precisión de las respuestas generadas, el sesgo potencial en los modelos de IA debido a los datos de entrenamiento, la posible reducción del pensamiento crítico…

Con respecto a los sesgos, en el artículo redactado por Partha Pratim Ray publicado en 2023 se menciona que estos sesgos se deben a los datos de entrenamiento, falta de conocimiento completo o actualizado, incapacidad para discernir la precisión factual, falta de conciencia contextual, limitaciones en el razonamiento ético y moral, desafíos con contextos conversacionales largos, incapacidad para generar contenido visual, dificultad para manejar solicitudes inapropiadas o dañinas, entre otros, de la inteligencia artificial.

Pistilli (2022), habla además de los debates éticos que plantean los *AGI (Artificial General Intelligence),* inteligencias artificiales generales, que define como sistemas de IA que se vuelven cada vez más especializados en tareas concretas, particularmente en el procesamiento del lenguaje natural.

En el estudio de Cazzaniga et al. (2024) se explica el efecto de la inteligencia artificial en el entorno laboral. En él se presenta cómo aproximadamente el 40% de los trabajadores a nivel mundial y el 60% en las economías avanzadas están en ocupaciones con alta exposición a la IA. El término exposición en su estudio se emplea para describir el grado de superposición entre las aplicaciones de inteligencia artificial (IA) y las habilidades humanas necesarias para realizar tareas en distintas ocupaciones. En otras palabras, se refiere a la capacidad de la IA para replicar la tarea de un determinado puesto de trabajo. En puestos con alta exposición existe mayor probabilidad de que esto suceda.

Asimismo, el estudio concluyó que las mujeres y los trabajadores con mayor nivel educativo tienen una mayor exposición a la IA, aunque esta es mitigada por el potencial más alto de complementariedad con la IA que tienen estos grupos. La capacidad de complementariedad se refiere a trabajar junto con la IA para aumentar la productividad, mejorar los resultados o realizar tareas más avanzadas que la IA no puede realizar por sí sola.

No pasa desapercibida tampoco la difusión de información falsa. Un estudio significativo realizado por Horne et al. en 2019, que investigó por primera vez cómo la asistencia de la inteligencia artificial afecta la percepción de la credibilidad de las noticias, lo evidenció. Los resultados revelaron que la asistencia de IA, especialmente cuando se proporcionaban explicaciones basadas en características, mejoraba significativamente la percepción de la confiabilidad y el sesgo de los artículos por parte de los consumidores de noticias. Este hallazgo lleva a considerar la posibilidad de utilizar inteligencias artificiales para abordar estos mismos problemas.

La privacidad es otro tema de preocupación. Elliott y Soifer (2022) en su estudio *«AI Technologies, Privacy, and Security»* (Tecnologías IA, Privacidad y seguridad) hablan sobre ello. Se menciona la violación del privilegio epistémico, que permite a las personas controlar qué aspectos de su identidad revelan o ocultan. Esto es debido a la vigilancia masiva, automatizada e indiscriminada, recopilando grandes cantidades de datos de personas a través de dispositivos cotidianos como teléfonos. Plantean que los sistemas de IA, al procesar y analizar esta información, suelen tener más conocimiento sobre las personas que ellas mismas, lo que trae preocupaciones sobre la privacidad y el desequilibrio en el control de la información personal.

3.3.2. La IA y su uso en educación y academia

La inteligencia artificial (IA) ha progresado bastante durante los últimos años ocasionando ciertos impactos en la educación, concretamente ChatGPT. Tras examinar cinco millones de estudios científicos

publicados el año pasado, el bibliotecario Andrew Gray detectó un aumento significativo en el uso de ciertas palabras en inglés, como «meticulosamente» (un 137% más), «intrincado» (117%), «encomiable» (83%) y «meticuloso» (59%). Gray atribuyó este fenómeno al uso de ChatGPT y programas similares de generación de lenguaje con IA para redactar o «pulir» estudios. Además, estimó que al menos 60,000 estudios científicos (más del 1% de los analizados en 2023) fueron escritos con la ayuda de esta herramienta (Ansede, 2024). Esto refleja el impacto de la IA en el lenguaje y la publicación de estudios científicos.

Paralelamente, es posible deducir que, al igual que los investigadores utilizan herramientas de IA para mejorar la calidad y el estilo de sus publicaciones, los estudiantes también podrían comenzar a depender de estas tecnologías para redactar. Esto podría llevar a una dependencia en el uso de estos modelos para escribir o reflejar un nivel que realmente no tienen. Nivel que se vería negativamente afectado al verse privados de esta herramienta. Tal dependencia podría generar desigualdades o crear falsos estándares entre los estudiantes. Asimismo, para los profesores, evaluar la originalidad y el pensamiento crítico del alumnado podría suponer un reto. En su artículo «La mentira de la IA», publicado en 2024, David Palumbo sostiene que la inteligencia artificial es perjudicial para el proceso creativo, ya que deteriora la calidad del trabajo, lo hace menos interesante y reduce la empleabilidad. Esta opinión, se basa en su experiencia como artista y profesional del arte comercial, habiendo trabajado como *freelancer*, director de arte y docente. No obstante, no todo son desventajas en el uso de la IA generativa. El Periodicó publicó en 2024 un artículo en el que se confirma que el equipo docente hace uso desde 2023 de una IA llamada Megaprofe para organizar la manera en que se dan las asignaturas con el objetivo de impartir clases ajustadas a las necesidades de cada alumno. Esta herramienta funciona de manera muy similar a ChatGPT ya que es un chatbot. A diferencia de ChatGPT, Megaprofe se mide con más parámetros, lo que permite al profesorado personalizar más la información según el caso dado. Expertos afirman que esta forma de emplear la IA trae consigo varios beneficios para los profesores como actualizar tablas de calendarios para ejercicios durante el año o transcribir un audio de una clase. El *CEO* de Megaprofe justifica su creación de la siguiente manera: «aunque nosotros no lo veamos», los maestros tienen «muchas cargas administrativas». Por ello, está plataforma está pensada para que la organización de las explicaciones sean más eficientes y menos repetitivas. Asimismo el *CEO* afirma que al usar esta IA se notará la reducción de horas que se emplean para llevar a cabo ciertas tareas (Ayuso. A, 2024).

En la facultad de derecho de la Universidad de Minnesota se le realizó una serie de exámenes a ChatGPT y se determinó que es capaz de adquirir un título universitario debido a su capacidad de generar respuestas de manera congruente (Lo, 2023) y referencias citadas en él. No obstante, el uso de esta IA genera tanto efectos positivos como negativos. Por un lado, las habilidades que demuestra al proporcionar respuestas son beneficiosas en el ámbito académico. Por otro lado, ChatGPT genera ciertas preocupaciones, como el posible impacto en el éxito académico de los alumnos si se depende exclusivamente de ella.

Por ello investigadores como Mhlanga o profesores como Sallam destacan la necesidad de un uso ético y responsable de esta IA, así como la preocupación por las respuestas erróneas que esta pueda generar (Lo, 2023). Además, en el estudio de la Universidad de Educación de Hong Kong realizado en 2023 y titulado «What is the impact of ChatGPT on education? A rapid review of the literature» (¿Cuál es el impacto de ChatGPT en la educación? Una revisión rápida de la literatura) se eligieron sistemáticamente 50 artículos académicos en inglés relacionados con la implicación de ChatGPT en la educación. Posteriormente, analizaron las respuestas que proporcionó ChatGPT 3.5 en estos artículos para evaluarlo y como resultado se obtuvo que ChatGPT dio resultados muy buenos en asignaturas como Economía y Pensamiento Crítico y de orden superior. No obstante, los resultados fueron insatisfactorios en asignaturas como Matemáticas y Educación médica. Finalmente, se llegó a la conclusión de que esta herramienta puede mejorar la enseñanza y el aprendizaje asistiendo tanto a los alumnos como a los profesores. Sin embargo, presenta ciertos dilemas como el plagio en la educación y la información inacabada o de dudosa veracidad que esta pueda llegar a brindar (Lo, 2023).

Teniendo en cuenta que este estudió se realizó con artículos exclusivamente en inglés, se podría considerar que los resultados obtenidos pueden variar examinado a ChatGPT en castellano.

Tabla 2

Evaluación del desempeño de ChatGPT-3.5 en diferentes ámbitos (de mejor a peor)

Ámbito	Desempeño global	Comentarios relevantes de los investigadores
Pensamiento crítico y habilidades de nivel superior	Sobresaliente	«Las respuestas generadas pueden ser evaluadas como claras en su exposición, precisas en cuanto a los ejemplos utilizados, y relevantes para las solicitudes...»
Economía	Sobresaliente	En TUCE, «ChatGPT se ubica en el percentil 99 en macroeconomía y en el percentil 91 en microeconomía, en comparación con (otros) estudiantes»
Programación	Sobresaliente a satisfactorio	«Casi todas las respuestas de ChatGPT fueron totalmente correctas y bien explicadas... las explicaciones proporcionadas por ChatGPT fueron sorprendentemente claras y orientadas hacia el objetivo» «El profesor calificó la tarea con una puntuación total de 71 sobre 100 puntos, lo que resultó en una calificación de 'Satisfactorio'»
Comprensión en Inglés	Satisfactorio	«La calificación promedio obtenida por ChatGPT según las normas oficiales fue de 7.18, lo cual es similar a la calificación promedio de todos los estudiantes en los Países Bajos»
Leyes	Apenas satisfactorio a insatisfactorio	«ChatGPT tuvo un desempeño en promedio al nivel de un estudiante de C+, logrando una calificación baja pero de aprobado» «Obtuvo su peor desempeño en exámenes que presentaban preguntas del estilo de problemas o de identificación de problemas»
Medicina	Apenas satisfactorio a insatisfactorio	«ChatGPT muestra una precisión moderada que se acerca al rendimiento de un aprobado en el *USMLE* (*United States Medical Licensing Exam* o Examen de licencia médica de los Estados Unidos)» «En los exámenes de BLS (Basic Life Support o soporte básico de vida) y ACLS (Advanced Cardiovascular Life Support o Apoyo Avanzado de Vida Cardiovascular) de la Asociación Americana del Corazón, ChatGPT no alcanzó el umbral de aprobado para ninguno de los exámenes»
Matemáticas	Insatisfactorio	«Las habilidades matemáticas de ChatGPT están significativamente por debajo del promedio de las de un estudiante de posgrado en matemáticas»
Evaluación de software	Insatisfactorio	«ChatGPT no es capaz, por sí solo, de aprobar un curso de evaluación de software. En total, ChatGPT pudo responder correctamente el 37.5% de las preguntas que planteamos»
Ciencias del deporte y psicología	Insatisfactorio	«ChatGPT respondió correctamente a varias de las 20 preguntas, pero suspendió el examen con una puntuación del 45%»
Exámenes de opción múltiple en diversas materias	Insatisfactorio	«ChatGPT no logra alcanzar la calificación de aprobado en casi ninguno de los exámenes de opción múltiple a los que se somete, y tiene un rendimiento significativamente inferior que el de un estudiante humano promedio»

Nota. Adaptado de «What is the impact of ChatGPT on education? A rapid review of the literature», por Lo, C. K., 2023, Education Sciences, 13(4), 410. p. 6. CC BY.

3.4. Consumo energético y sostenibilidad de la IA

El efecto ambiental generado por los recursos computacionales requeridos para entrenar y ejecutar los modelos de ChatGPT y otras IA generativas, abarcando tanto el consumo energético y de agua, como las emisiones de carbono es una preocupación creciente y cuyos efectos sería importante considerar.

Por un lado, GPT-3 cuenta con 10 veces más parámetros que cualquier otro modelo de lenguaje no disperso disponible (Patterson et al., 2021). Se calcula que sus emisiones de carbono durante el entrenamiento son de 552 toneladas de CO_2 equivalente (tCO_2e) y que consume 1287 megavatios hora (MWh) de energía (Patterson et al., 2021).

El Gobierno de Navarra consume un total de 154.618.544 kWh/año, lo que equivale al consumo de 12884 hogares según datos del Sistema de Información Energética (SIE) en octubre de 2024. Teniendo estos datos en cuenta, el consumo de energía necesario para el entrenamiento de ChatGPT equivale al de aproximadamente 107 hogares navarros en un año. En consecuencia, la sostenibilidad de las IA *'s* plantea ciertas dificultades para alcanzar los Objetivos de Desarrollo Sostenible (ODS) que fueron establecidas en la Agenda para el Desarrollo Sostenible en relación al cambio climático, en concreto el ODS 13. (Ayala, 2021). No obstante, debido a la gran expansión de la IA y su impacto en la sociedad, la Declaración de Montreal sobre la sostenibilidad en la era digital expuso en 2020 la significate ayuda que esta herramienta supondría para llegar a transformaciones sociales, que a su vez puedan reducir las emisiones de CO_2, llegando así a los Objetivos de Desarrollo Sostenible. Asimismo, se llevó a cabo una investigación para determinar si el uso de la IA resultaba beneficioso o perjudicial en relación a los ODS. Con ello se concluyó que la IA puede conseguir alcanzar 134 objetivos, pero a su vez impedir que se lleguen a cumplir 59 propósitos. En conclusión la IA en el ámbito climático puede ser beneficiosa, ayudando a la perfección de los recursos naturales y reducir el impacto medioambiental ocasionada por las actividades humanas, entre otras… o bien, ser perjudiciales como el alto consumo energético que se requiere y que bien se ha mencionado antes o dejando rastros de carbono en el medioambiente. (ACUÑA, 2024)

Tabla 3

Relación entre el consumo de energía de un hogar Navarro con el consumo en el entrenamiento del Modelo ChatGPT según datos del SIE y el estudio de Patterson et al., 2021

Kilovatios hora (KWh) que consume ChatGPT durante su entrenamiento	Kilovatios hora que consume aproximadamente un hogar Navarro en un año	Relación
1287000	12000	107

Nota. Tabla de elaboración propia en la que se muestra la cantidad de energía que consume ChatGPT.

Además, se estima también que una búsqueda realizada por IA generativa utiliza entre cuatro y cinco veces más energía que una búsqueda web convencional. (Crawford, 2024). En particular, se estima que el entrenamiento de ChatGPT-3 consumió 78437 kWh de electricidad, una cantidad equivalente a la energía utilizada en una casa en España durante 23 años (Europa Press, 2023).

Los centros de datos que entrenan modelos de IA, como GPT-3 y GPT-4, representan alrededor del 1-2% del uso global de electricidad en todo el planeta, lo que supone un desastre ambiental. Esta alta demanda de energía contribuye significativamente a la huella de carbono de la tecnología (Li et al., 2023).

Por otro lado, el agua es un recurso crucial en la producción de IA, utilizado para enfriar los procesadores y generar electricidad. Para enfriar los servidores de ChatGPT-3 se utilizaron 70,000 litros de agua dulce limpia (Europa Press, 2023). Esto es equivalente al agua utilizada por más de 547 personas en España durante un día, considerando que el consumo medio de agua por habitante en España es de 128 litros al día en 2022 según datos del Instituto Nacional de Estadística (INE). El uso de ChatGPT también implica

un consumo de agua: una conversación con ChatGPT de entre 20 y 50 preguntas gasta aproximadamente 0.5 litros (Europa Press, 2023).

El West Des Moines, Iowa, un gran centro de datos que apoya al modelo GPT-4 de OpenAI utilizó aproximadamente el 6% del suministro de agua del distrito en julio de 2022, justo antes de que OpenAI completara el entrenamiento del modelo (Crawford, 2024).

Google y Microsoft, al desarrollar sus modelos de lenguaje Bard y Bing, vieron incrementos notables en su consumo de agua, con aumentos del 20% y 34% respectivamente en el transcurso de un año (Crawford, 2024). La combinación del uso de agua en los centros de datos de Google, Microsoft y Meta alcanzó aproximadamente 2.2 mil millones de metros cúbicos en 2022 (Crawford, 2024). Esto es equivalente al uso anual total de agua de dos Dinamarcas. A nivel global, se proyecta que la demanda de agua para la inteligencia artificial podría alcanzar una cifra equivalente a la mitad del consumo total del Reino Unido para el año 2027 (Crawford, 2024).

Tabla 4

Estimación de la huella promedio de consumo operativo de agua de GPT-3. ‘*’ *indica centros de datos en construcción a partir de julio 2023, y los valores PUE y WUE para estos centros de datos se basan en la proyección de Microsoft.*

Ubicación	PUE	WUE (L/kWh)	Intensidad de agua para electricidad (L/kWh)	Agua TOTAL para entrenamiento (millones de L)	Agua TOTAL por cada inferencia (mL)	# Número de inferencias por 500 mL de agua
Promedio en EE.UU.	1,17	0,55	3,142	5,439	16,904	29,6

Nota: Adaptado de «*Making AI Less 'Thirsty': Uncovering and Addressing the Secret Water Footprint of AI Models*», por P. Li, J. Yang, M. A. Islam, y S. Ren, 2023, *arXiv preprint arXiv:2304.03271*, p.

7. CC BY-NC

En la tabla 4, se pueden ver el PUE y el WUE de los diferentes centros de datos del modelo GPT-3 Además se observa la cantidad de agua empleada para entrenar al modelo y por interacción.

La Eficiencia del Uso de Energía (PUE) es una métrica estándar utilizada para medir la eficiencia energética en los centros de datos. Se obtiene dividiendo la energía total consumida por el centro de datos entre la energía utilizada específicamente por el equipo de TI. Un PUE cercano a 1 indica que la mayor parte de la energía se destina al funcionamiento del equipo de TI, lo que sugiere una alta eficiencia energética. No obstante, el PUE solo mide la eficiencia en términos energéticos y no refleja completamente el impacto ambiental global del centro de datos (Carrillo M-Feduchi, 2022).

La Eficiencia del Uso del Agua (WUE), definida por The Green Grid, evalúa la cantidad de agua que consume un centro de datos en relación con la energía utilizada por sus equipos de Tecnologías de la Información (TI). Se calcula dividiendo el consumo anual de agua, en litros, por la energía empleada por los equipos de TI, medida en kilovatios-hora (kWh). Este índice considera el agua utilizada para enfriamiento, control de humedad y producción de electricidad en el sitio. El WUE tiene como objetivo medir el impacto del consumo de agua en la eficiencia global del centro de datos y contribuir a la reducción de su impacto ambiental en relación con el agua (Carrillo M-Feduchi, 2022).

La Eficiencia del Uso del Carbono (CUE) evalúa las emisiones de carbono asociadas con un centro de datos, ayudando a las organizaciones a entender y reducir su huella de carbono. Un valor ideal de CUE es 0.0, lo que indicaría la ausencia de emisiones de carbono en las operaciones del centro de datos. A diferencia del PUE, que busca un valor cercano a 1.0, el CUE se enfoca en minimizar el impacto ambiental relacionado

con el carbono y puede facilitar la obtención de incentivos por prácticas de computación verde (Carrillo M-Feduchi, 2022).

Como se observa, en la tabla, si se realiza la media entre todos los centros de agua empleada por interacción, se obtiene un valor de 19mL. Lo cual puede parecer insignificante, pero si se tiene en cuenta el número de interacciones que un solo usuario puede hacer por día y el número de usuarios total que pueden tener este tipo de inteligencias de chat generativo, las cifras son preocupantes.

Para abordar estos desafíos, el 1 de febrero de 2024, un grupo de senadores y representantes en EE. UU., como Edward J. Markey y Martin Heinrich presentó un proyecto de ley destinado a abordar los impactos ambientales de la inteligencia artificial. En él, se propone que el Instituto Nacional de Estándares y Tecnología (NIST) desarrolle estándares para medir y reportar los impactos ambientales de la IA, así como un marco voluntario para que los desarrolladores de IA informen sobre estos impactos. Además, se requerirá un estudio interagencial por parte de la Agencia de Protección Ambiental (EPA) sobre los efectos positivos y negativos de la IA en el medio ambiente, abarcando desde el consumo energético hasta los residuos electrónicos (Markey et al., 2024).

4. Objetivos e hipótesis

En el presente trabajo se ha planteado la consecución de los siguientes objetivos:

1. Realizar un formulario y extraer conclusiones sobre el uso y percepción de la herramienta generativa ChatGPT de docentes y estudiantes sobre la herramienta.
2. Someter a ChatGPT-3.5, ChatGPT-4, y ChatGPT-4o a exámenes de filosofía y a ChatGPT-3.5, ChatGPT-4, ChatGPT-4o, ChatGPT-4o en You.com y ChatGPT-o1 preview a exámenes de física.
3. Evaluar y analizar el desempeño de los modelos mediante una serie de rúbricas de creación propia.
4. Comparar el desempeño de los modelos entre sí.

Teniendo en cuenta dichos objetivos planteamos las siguientes hipótesis:

1. Con respecto al análisis de las respuestas a los cuestionarios, se espera:
   - 1.1. Que los docentes la usen significativamente más que el alumnado.
   - 1.2. Que en el cuerpo docente se usen más las herramientas generativas de lenguaje entre aquellos con menos años de experiencia.
   - 1.3. Que tanto en el alumnado como en el profesorado haya una distribución uniforme de las respuestas asociadas a la variable género.
   - 1.4. Que en el alumnado la edad covaríe positivamente con el uso de las herramientas que estamos estudiando, entre el profesorado se espera la tendencia contraria.
   - 1.5. Que la percepción de los estudiantes en el fallo de las respuestas sea mayor en las asignaturas de matemáticas y física.
   - 1.6. Que los alumnos prefieran el apoyo de un profesor que el de la IA.
2. La eficacia del modelo Chat GPT-3.5 en la generación de respuestas atendiendo al número de revisiones del *prompt* será significativamente menor en preguntas de filosofía que en los problemas de física.
3. Se espera una mejora significativa en las respuestas de la herramienta en sus versiones más recientes y de pago, especialmente en la resolución de problemas de física.

5. Metodología

5.1. Material

Para realizar este estudio exploratorio y a su vez cualitativo por una parte y cuantitativo por otra, se necesitarán diversos materiales y recursos:

5.1.1. Obtención de datos

Los datos se obtienen por diversas vías:

a) Dos cuestionarios enfocados en el uso y percepción de la herramienta generativa ChatGPT. Dirigidos a estudiantes y profesores respectivamente.
b) En segundo lugar directamente de ChatGPT al pedirle que resuelva bien una serie de exámenes de Bachillerato de filosofía y de física que se corrigen con una rúbrica elaborada para evaluar las funciones de ChatGPT.

5.1.2. ChatGPT

Se hará uso de una interfaz de programación de aplicaciones de OpenAI, conocida como *API* (*Application Programming Interface)*. *API* es un conjunto de definiciones y protocolos que les permiten a dos componentes de software comunicarse entre sí (Rivera, 2023). Se empleará para acceder a una instancia de ChatGPT y las versiones de pago de ChatGPT. Asimismo, se hará uso de un buscador alternativo desarrollado por una IA conocida como *You.com*.

5.1.3. Herramientas de análisis y visualización

Para analizar los resultados de los experimentos se emplea excel o su versión *googlesheets* y en caso de requerir gráficos o análisis estadísticos más avanzados spss. En el caso de las tablas especialmente en la discusión, que aparecen sin fuente, son de elaboración propia.

5.2. Procedimiento

5.2.1. Parte 1: Cuestionarios sobre el uso y percepción de ChatGPT dirigido a estudiantes y docentes.

Se elabora un formulario de Google específico para alumnos y otro para docentes. Se recogen las respuestas y se analizan los resultados. Los formularios (Anexo A) realizan preguntas sobre la frecuencia, hábitos y percepción relacionadas al uso de ChatGPT.

5.2.2. Parte 2: Modelos de examen con ChatGPT-3.5, ChatGPT-4, ChatGPT-4o, y ChatGPT o1-preview.

Se recoge una serie de exámenes de la asignatura de filosofía y la de física (Anexo B). El primer exámen de filosofía pertenece al temario de 1º de Bachiller concretamente es sobre el surgimiento de la filosofía. La segunda prueba de filosofía consiste en la realización de una disertación a partir de un texto sobre el ocio, también del temario de 1º de Bachiller. Todas las pruebas de filosofía mencionadas anteriormente se realizan a los modelos ChatGPT-3.5, ChatGPT-4 y ChatGPT-4o. Sin embargo, a partir de la tercera prueba solo es posible realizarlo al modelo ChatGPT-4 y ChatGPT-4o debido a la inhabilitación del primero a partir del 19 de Julio de 2024 (AndroidSage, 2024). Se realizan otros 6 exámenes de filosofía pertenecientes a la EVAU de Madrid en los años 2023 y 2021.

Una vez finalizado con las pruebas de filosofía sometemos también a la IA a las pruebas de física. Realizamos un total de 10 problemas del nivel de 3º ESO y 10 de 2º de Bachiller al modelo. Los primeros 10 se realizan a ChatGPT-4, ChatGPT-4o, ChatGPT-4o en You.com y ChatGPT-o1. Los últimos 10 solamente a ChatGPT-4 y ChatGPT-4o debido a limitaciones de tiempo. El primer problema de 3º ESO es el único que se logra realizar a ChatGPT-3.5 antes del cierre de este modelo.

Tras recoger las respuestas, se evalúan siguiendo una serie de rúbricas de elaboración propia, personalizadas y adaptadas, en cada una de las pruebas de filosofía (especificando el contenido mínimo) y utilizado una común para las de física. En los de física se evalúan aspectos como el número de interacciones necesarias para una respuesta satisfactoria, el número de errores cometidos, así como el número de intentos necesarios para llevar a cabo la resolución del problema y la explicación del proceso. En los de filosofía se

tiene en cuenta la claridad y la profundidad del contenido o la precisión del vocabulario. También, en la disertación la capacidad de presentar argumentos propios.

5.2.3. Comparación de la Performance de los modelos.

Finalmente se hace una comparación del desempeño *(performance)* entre cada uno de los modelos tomando la nota media o global.

En resumen, tras realizar la experimentación se obtiene una reflexión sobre el funcionamiento y desempeño de estos modelos comprendiendo sus puntos fuertes y débiles. Se observa en que fallan y que realizan bien y se visualiza si existe una mejora real y efectiva en el desempeño de estos en las pruebas según son más recientes.

6. Datos

6.1. Análisis estadístico del uso de Chat GPT en educación.

Una vez realizados los cuestionarios de investigación sobre el impacto de la inteligencia artificial (IA) en la educación dirigidos al alumnado y profesorado del IES Plaza de la Cruz, se cuenta con una muestra total de 175 respuestas de las cuales 66 son del profesorado y 109 del alumnado. En el Anexo I se recogen las respuestas correspondientes a cada formulario en cada pregunta así como diferentes gráficas analizando las respuestas.

Tabla 5

Distribución de las variables recogidas en la muestra del IES Plaza de la Cruz

	N	Mínimo	Máximo	Media	Desv. estándar	Varianza
USO	175	1	5	2,47	0,999	0,998
Fiabilidad	175	1	4	2,81	0,860	0,740
Edad (años)	175	12	62	27,26	16,503	272,333
Experiencia	66	1	4	2,79	1,183	1,400

Nota. Esta tabla representa las variables de uso, fiabilidad, edad y la experiencia del uso de IA que han tenido los encuestados de Plaza de la Cruz.

6.2. Estadísticos descriptivos de las variables cualitativas

Tabla 6

Categoría de los encuestados

	Categoría			
	Frecuencia	**Porcentaje**	**Porcentaje válido**	**Porcentaje acumulado**
Profesorado	66	37,7	37,7	37,7
Alumnado	109	62,3	62,3	100,0
Total	175	100	100	

Tabla 7

Frecuencia de cada categoría de la variable de género

Categoría				
	Frecuencia	**Porcentaje**	**Porcentaje válido**	**Porcentaje acumulado**
Masculino	63	36	36	36
Femenino	108	61,7	61,7	97,7
Otro	4	2,3	2,3	100,0
Total	175	100	100	

Tabla 8

Uso de la herramienta en contexto educativo

Categoría				
	Frecuencia	**Porcentaje**	**Porcentaje válido**	**Porcentaje acumulado**
Nunca	29	16,6	16,6	16,6
Raramente	67	38,3	38,3	54,9
Ocasionalmente	51	36,6	36,6	84,0
Frecuentemente	24	29,2	29,2	97,7
A diario	4	13,7	13,7	100,0
Total	175	100	100	

Tabla 9

Fiabilidad de la herramienta según la percepción del usuario

Categoría				
	Frecuencia	**Porcentaje**	**Porcentaje válido**	**Porcentaje acumulado**
Poco fiable	8	4,6	4,6	4,6
Medianamente fiable	60	34,3	34,3	38,9
Bastante fiable	64	36,6	36,6	75,4
No se conoce	43	24,6	24,6	100,0
Total	175	100	100	

Tabla 10

Años de experiencia docente. (Solo aplica a la muestra de profesorado, población 66)

Categoría				
	Frecuencia	**Porcentaje**	**Porcentaje válido**	**Porcentaje acumulado**

de 0 a cinco años	14	8,0	21,2	21,2
de 5 a 10 años	12	6,9	18,2	39,4
de 10 a 20 años	14	8,0	21,2	60,6
más de 20 años	26	14,9	39,4	100,0
Total	66	37,7	100	
Sistema	109	62,3		

No se incluye en este la tabla de frecuencias de la variable edad, ya que las edades del profesorado presentan una gran dispersión entre los 23 y 62 años. Esta queda recogida en el Anexo A.

Tabla 11

Edad

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	12	1	,6	,6	,6
	13	26	14,9	14,9	15,4
	14	11	6,3	6,3	21,7
	15	11	6,3	6,3	20,0
	16	19	10,9	10,9	38,9
	17	32	18,3	18,3	57,1
	18	7	4,0	4,0	61,1
	19	2	1,1	1,1	62,3
	23	1	,6	,6	62,9
	26	1	,6	,6	63,4
	27	1	,6	,6	64,0
	29	3	1,7	1,7	65,7
	30	2	1,1	1,1	66,9

31	1	,6	,6	67,4
34	1	,6	,6	68,0
37	1	,6	,6	68,6
38	2	1,1	1,1	69,7
39	1	,6	,6	70,3
40	2	1,1	1,1	71,4
41	2	1,1	1,1	72,6
42	2	1,1	1,1	73,7
43	3	1,7	1,7	75,4
44	1	,6	,6	76,0
45	1	,6	,6	76,6
46	2	1,1	1,1	77,7
47	5	2,9	2,9	80,6
48	2	1,1	1,1	81,7
49	3	1,7	1,7	83,4
50	3	1,7	1,7	85,1
51	1	,6	,6	85,7
52	1	,6	,6	86,3
53	1	,6	,6	86,9
54	6	3,4	3,4	90,0
55	2	1,1	1,1	91,4
56	1	,6	,6	92,0

57	1	,6	,6	92,6
58	9	5,1	5,1	97,7
59	2	1,1	1,1	98,9
61	1	,6	,6	99,4
62	1	,6	,6	100,0
TOTAL	175	100,00	100,00	

Nota. Esta tabla representa muestra la edad de los encuestados, así como la cantidad de veces que se repiten las edades (frecuencia) y sus correspondientes porcentajes.

6.3. Evaluación de ChatGPT ante los exámenes mediante rúbricas.

Las respuestas del modelo se pueden ver en el Anexo B y la evaluación detallada en el Anexo C. A continuación, se reflejan algunas comparaciones entre los modelos tomando todos los exámenes comunes en los que se examinan a la vez.

Tabla 12

Comparación de los tres modelos en filosofía.

ChatGPT-3,5	4,75
ChatGPT-4	7,45
ChatGPT-4o	7,48

Tabla 13

Comparación de los dos modelos en filosofía.

ChatGPT-4	7,01
ChatGPT-4o	7,50

Tabla 14

Comparación de todos los modelos analizados en física.

ChatGPT-3,5	3,33
ChatGPT-4	10,00
ChatGPT-4o	10,00
ChatGPT-4o en You.com	9,17
ChatGPT-o1 preview	10,00

Tabla 15

Comparación entre los modelos más recientes en física.

ChatGPT-4	9,07
ChatGPT-4o	9,07
ChatGPT-4o en You.com	8,33
ChatGPT-o1 preview	8,75

Tabla 16

Comparación entre los dos modelos en física.

ChatGPT-4	8,80
ChatGPT-4o	8,63

7. Discusión de resultados

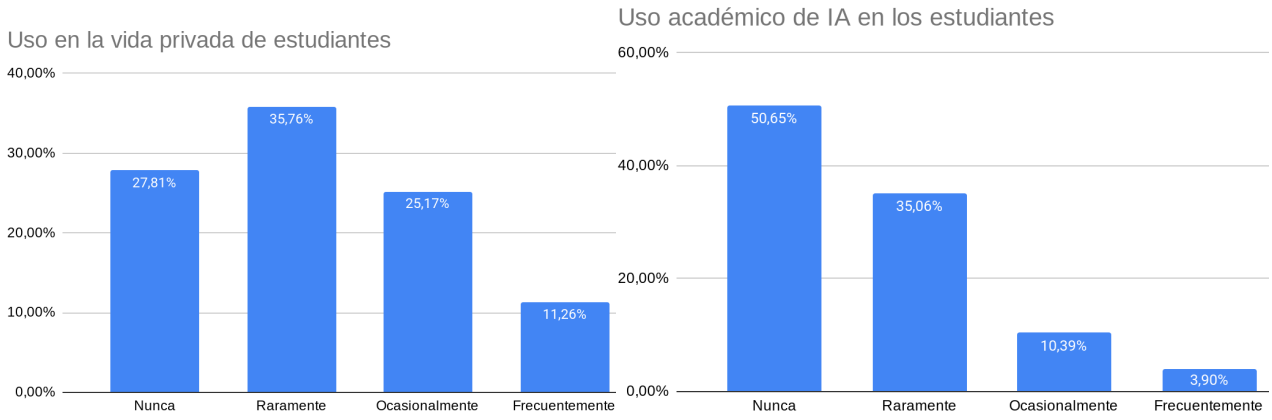
7.1. Discusión estadística de los resultados de las encuestas.

A continuación, se discuten los resultados más relevantes de las respuestas a los formularios. En el Anexo A se pueden observar más gráficas.

7.1.1. Formulario de investigación sobre el impacto de la inteligencia artificial (IA) en la educación dirigido específicamente a estudiantes.

Figura 5

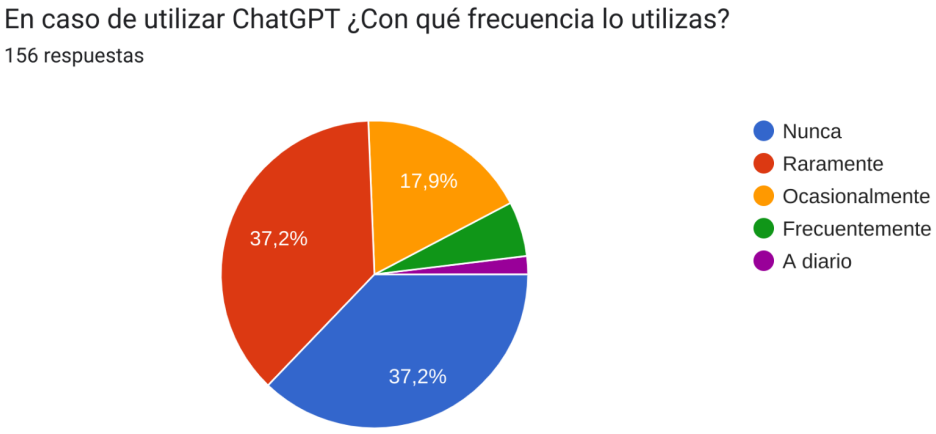
Gráficas de uso de los estudiantes en la vida privada y académica



En la figura 5, se observa que el reporte más habitual es emplear la IA raramente en la vida privada (con un 35,76% de las respuestas marcando esto), seguido de nunca (27,81%) y ocasionalmente (25,17%). En la vida académica lo más común es que no se utilice nunca, con un 50,65% de las respuestas, seguida de la misma tendencia relacionada a un poco uso, con solamente un 3,9% reportando que la usa frecuentemente.

Figura 6

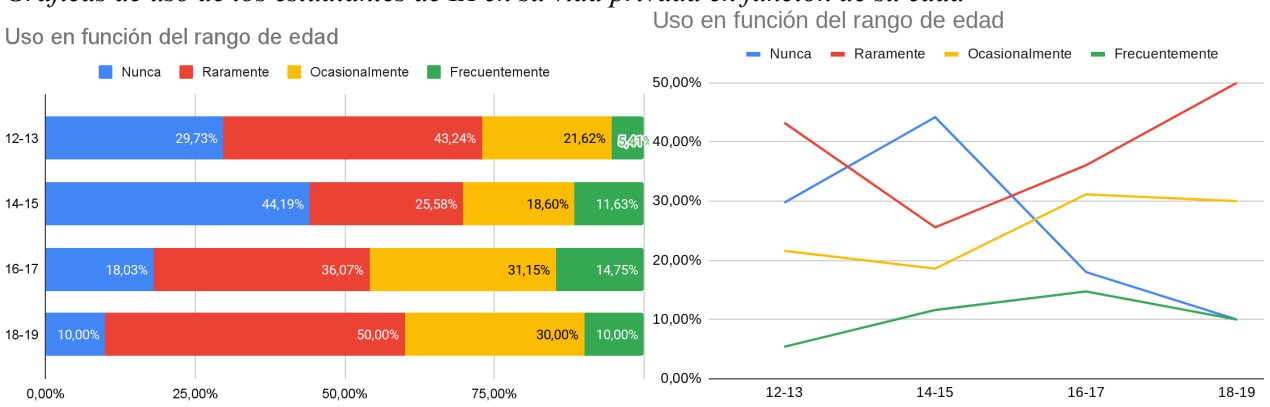
Gráfica de uso de los estudiantes de ChatGPT



Respecto al uso concreto de la IA ChatGPT, en la figura 6 se ve que las respuestas se igualan en un 37,2% en el uso raramente y nunca. Igualmente, la tendencia más común es una poca frecuencia de uso.

Figura 7

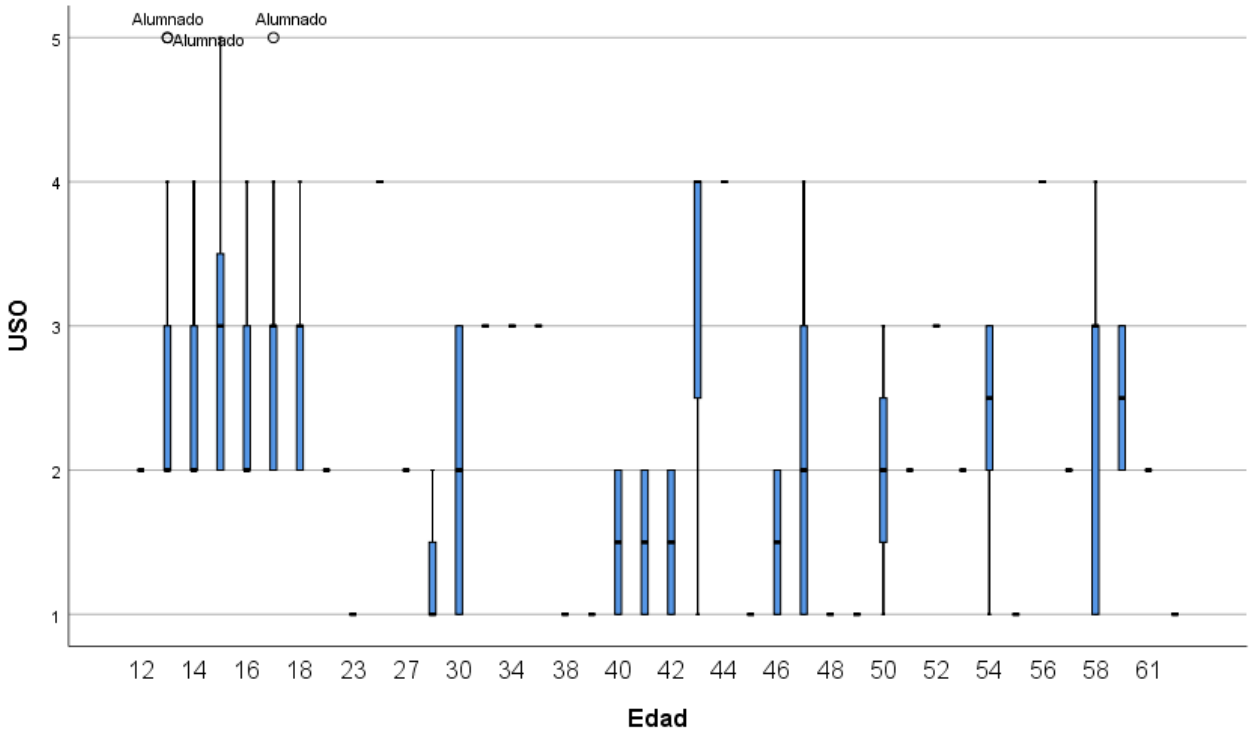
Gráficas de uso de los estudiantes de IA en su vida privada en función de su edad



En función a la variable edad, el uso «frecuentemente» es más común en el grupo de 16 a 17 años de edad con un reporte del 14,75%, tal y como se observa en la figura 7. La tendencia más común es usarla raramente, con el mayor número de reportes en cada grupo (un 43,24% frente al 29,73% de nunca, el 21,62% de ocasionalmente y el 5,41% de frecuentemente en el grupo de 12-13 años por ejemplo), excepto en el de 14 a 15 años, que es el grupo que menos reporta usarlo con la variable «nunca» en un 44,19%. Se observa como la variable nunca toma los menores valores en los dos grupos de mayor edad.

Figura 8

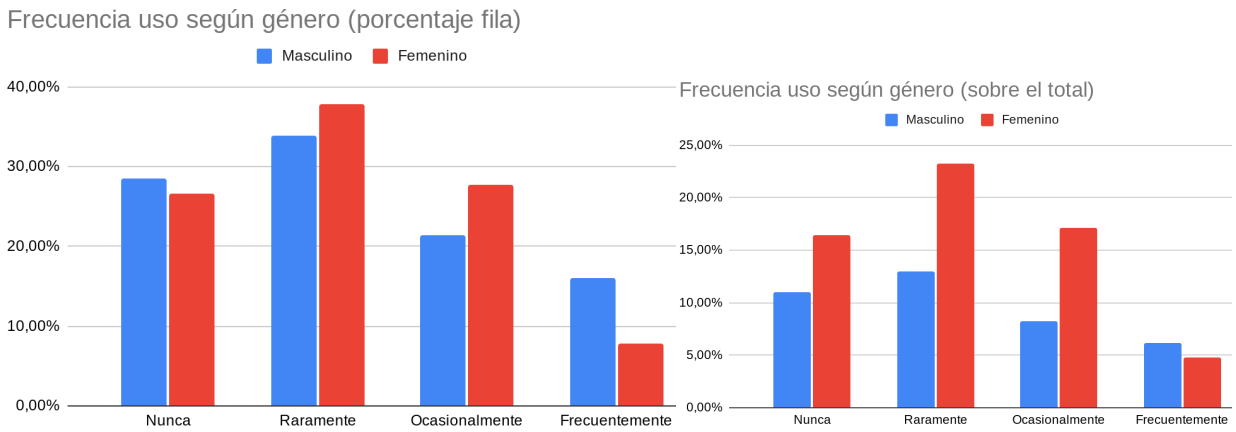
Análisis de barras y bigotes del uso de ChatGPT según la edad de los encuestados.



Aunque en general el alumnado hace mayor uso de la IA que el profesorado no se cumple la hipótesis cuatro de que el uso de la IA covaría positivamente con la edad entre el alumnado y negativamente entre el profesorado. Encontrando que no existen diferencias significativas entre la edad y el uso de la herramienta de inteligencia artificial.

Figura 9

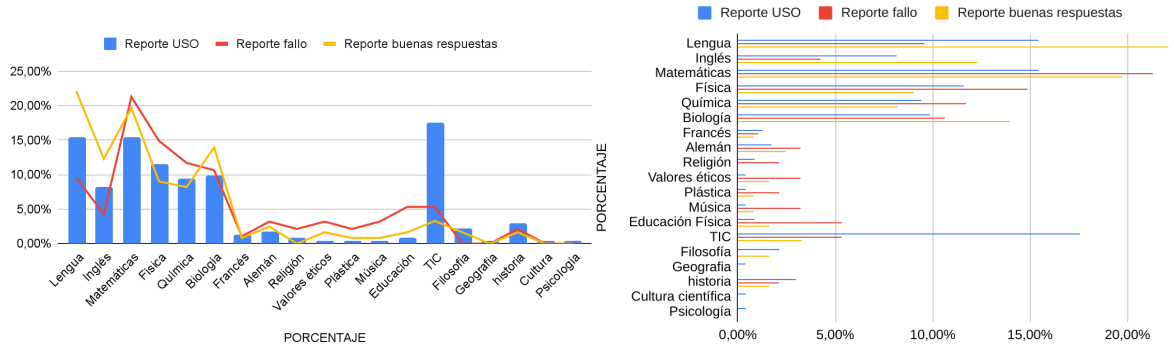
Gráficas de uso de los estudiantes de IA en su vida privada en función de su género



En la figura 9, se observa que no existe una variación muy alta en la frecuencia de uso según el género. Aunque el femenino reporta más usarlo ocasionalmente y raramente, el masculino reporta más usarlo frecuentemente. No se extraen conclusiones muy drásticas de estas.

Figura 10

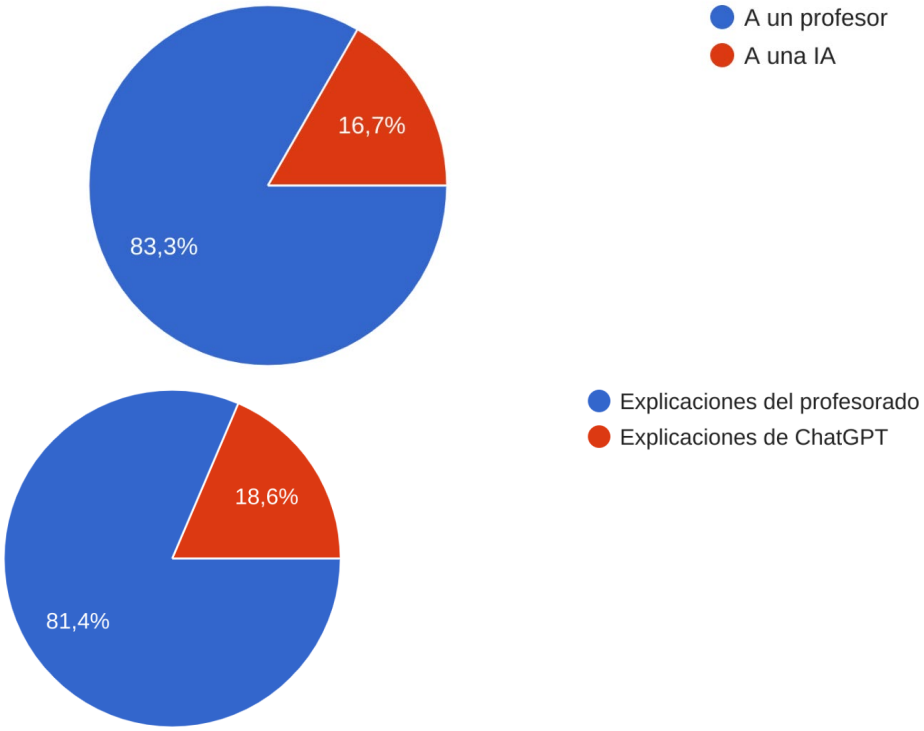
Gráficas comparativas entre el uso y el reporte de fallos o buenas respuestas



En las dos anteriores gráficas (Figura 10) se puede ver la relación entre el uso, el reporte de fallos y el de buenas respuestas. La asignatura en la que más uso se reporta es TIC, seguida de lengua y matemáticas. La asignatura en la que más fallos se reporta es matemáticas, seguida de física y biología. La asignatura en la que mejores respuestas se reporta es lengua, seguida de matemáticas e inglés. Por lo general el valor del uso de una asignatura es cercano al de su reporte de fallo y mejores respuestas. Aunque en casos como el de TIC se reportan mucho uso y poco fallo o mejores respuestas, siendo el primer caso algo más común. Esto corrobora nuestra hipótesis 5.

Figura 11

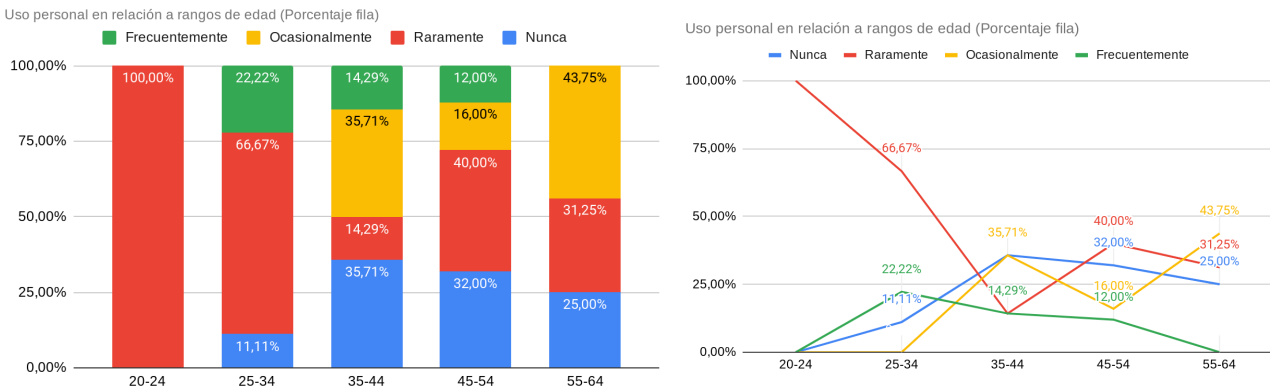
Gráficas sobre la preferencia a preguntar a un profesor o a una IA y la explicación de un profesor o de ChatGPT.



7.1.2. Formulario de investigación sobre el impacto de la inteligencia artificial (IA) en la educación dirigido específicamente a docentes.

Figura 12

Gráficas de uso de los docentes de IA en su vida privada en función de su edad por rango



Fuente: elaboración propia.

En la figura 12, el grupo que menos reporta usarlo mediante la variable nunca es el de 35 a 44 años de edad, seguido del de 45 a 54 años de edad. El que más reporta usarlo frecuentemente es el grupo de entre 25-34 años de edad. El grupo que más reporta usarlo ocasionalmente es el de 55 a 64 años de edad. Entre el grupo de 20 a 24 años todas las respuestas que se obtienen indican usarla raramente.

Para analizar si existe una tendencia en el uso según el rango de edad, se asigna un valor numérico a cada respuesta cualitativa: nunca = 0, raramente = 1, ocasionalmente = 2 y frecuentemente = 3. Posteriormente, se calcula la frecuencia de cada respuesta en cada rango de edad y se pondera utilizando estos valores numéricos. El resultado se observa en la tabla 17.

Se observa que los resultados no muestran una tendencia clara en el uso según la edad. Sin embargo, el rango de edad de 35-44 años presenta el puntaje más alto (1.50), lo que indica que este grupo tiende a usar la IA con mayor frecuencia en comparación con los demás. Por otro lado, aunque el grupo de 55-64 años no alcanza los niveles de 35-44, su puntaje (1.19) es superior al del grupo más joven (20-24, con 1.00).

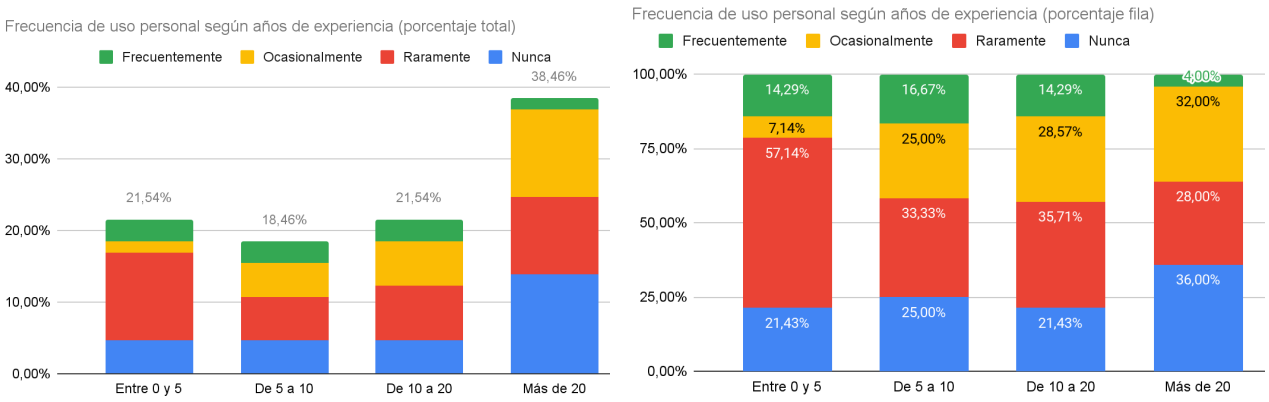
Tabla 17

Distribución de las variables recogidas en la muestra del IES Plaza de la Cruz

Rango de edad	Frecuencia de uso
20-24	1
25-34	1,33
35-44	1,5
45-54	1,16
55-64	1,19

Figura 13

Gráficas de uso de los docentes de IA en su vida privada en función de sus años de experiencia



Nota. En esta figura se observa que el grupo que más reporta usarla nunca es el de más de 20 años de experiencia. A su vez, es el grupo que menos reporta usarla frecuentemente y más ocasionalmente. La moda es usarla raramente, aunque se observa una ligera variación de esta frente al uso ocasional en el grupo de más años de experiencia. Los grupos que reportan usarla más frecuentemente con un porcentaje igualado de 14,29% son el de entre 0 y 5 años de experiencia y el de entre 10 a 20 años de experiencia.

Tabla 18

Pruebas de normalidad de la variable uso de la IA en relación con la experiencia docente.

Pruebas de normalidad							
		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Experiencia	Estadístico	gl	Sig.	Estadístico	gl	Sig.
USO	de 0 a 5 años	,255	14	,014	,828	14	,011
	de 5 a 10 años	,265	12	,020	,779	12	,005
	de 10 a 20 años	,263	14	,009	,806	14	,006
	más de 20 años	,309	26	,000	,782	26	,000

a. Corrección de significación de Lilliefors

La prueba de normalidad se lleva a cabo con el programa de cálculo estadístico spss, para cada una de las distribuciones. Si se cumplen con normalidad,esto es si la significancia asintótica bilateral obtenida en las pruebas de Shapiro-Wilk (para más de 50 sujetos de muestra) es <0.05 se considerará «no normal» y si el valor es >0.05 se considerará «normal». De esta manera podemos comparar con un test ANOVA ambas variables y así comprobar o refutar la hipótesis 2.

En la tabla 18 se puede observar como la significancia entre grupos es mayor de 0,05 lo que quiere decir que las diferencias entre el uso de Chat GPT en contexto educativo según los años de experiencia no son significativas para un intervalo de confianza del 95%. Lo cual refuta nuestra segunda hipótesis.

Tabla 19

Relación del uso de ChatGPT entre los distintos grupos cuestonados

ANOVA

USO

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Entre grupos	2,936	3	,979	,842	,476
Dentro de grupos	72,049	62	1,162		
Total	74,985	65			

Figura 14

Representación gráfica de la relación existente entre el uso de la Inteligencia artificial y los años de experiencia docente

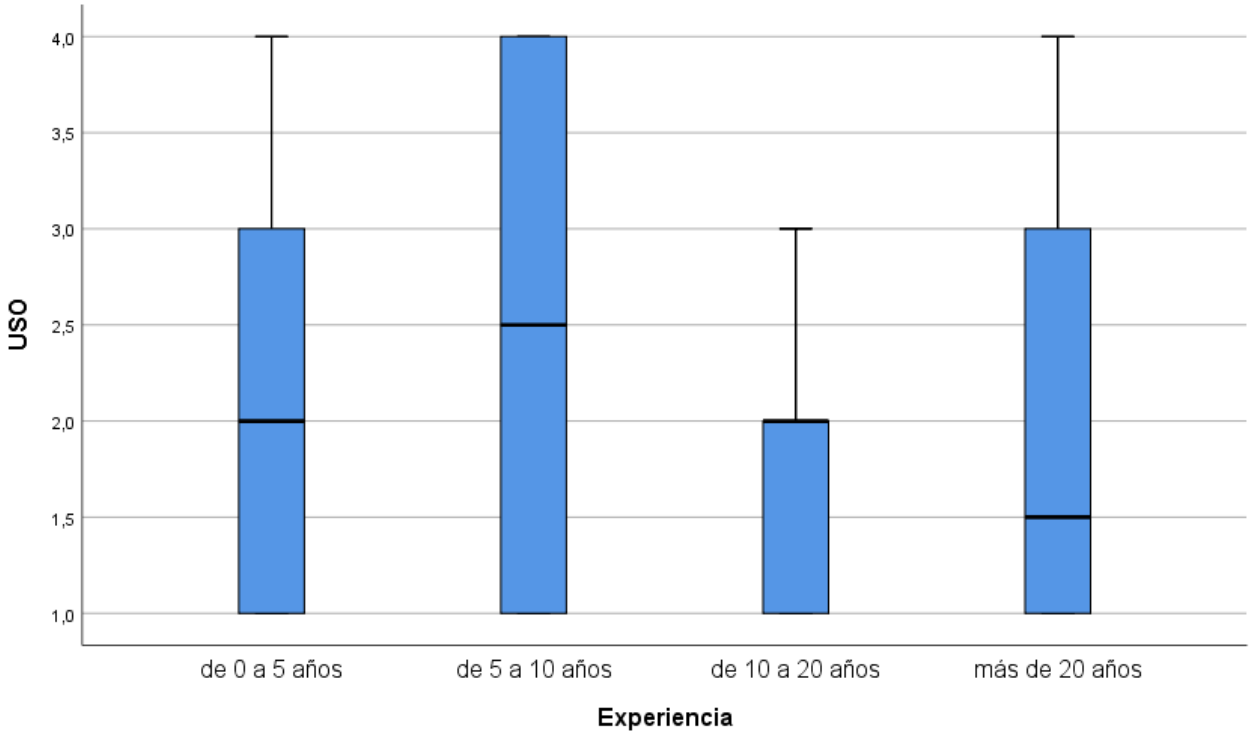


Figura 15

Gráfica de uso de los docentes de IA en su vida privada en función de su género

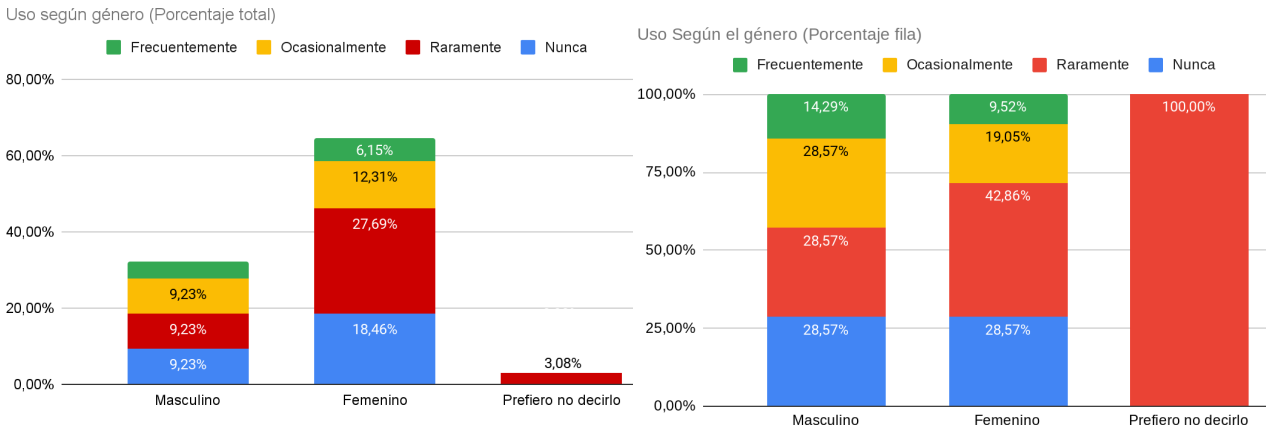


Figura 16

Gráfica de la disponibilidad para pagar por una versión premium de IA en función de la frecuencia de uso en la vida privada

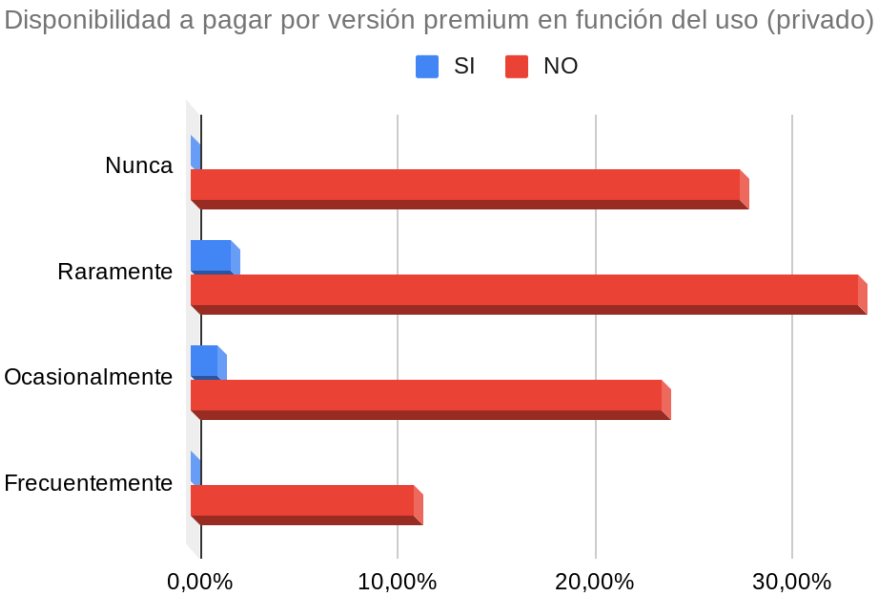
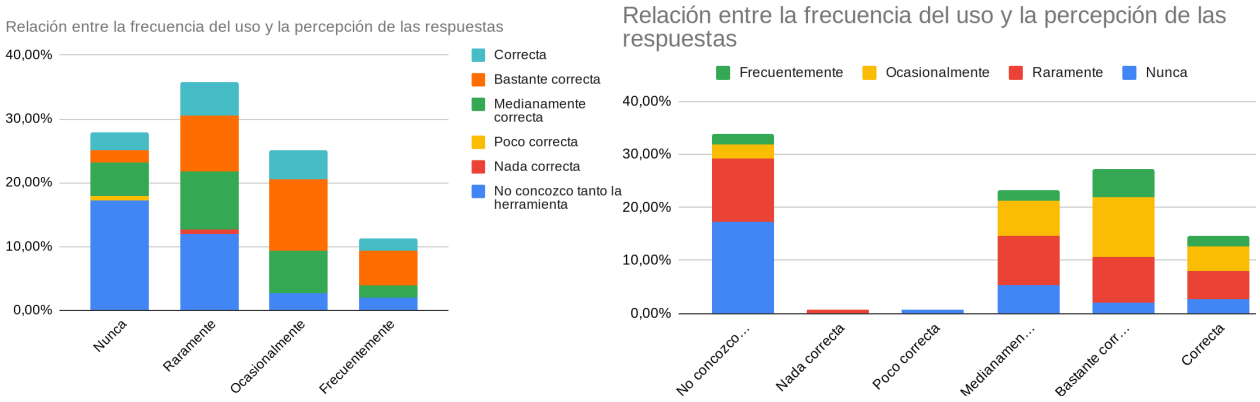


Figura 17

Gráficas sobre la percepción en la corrección de respuestas por los profesores en función a la frecuencia de uso de IA en su vida privada

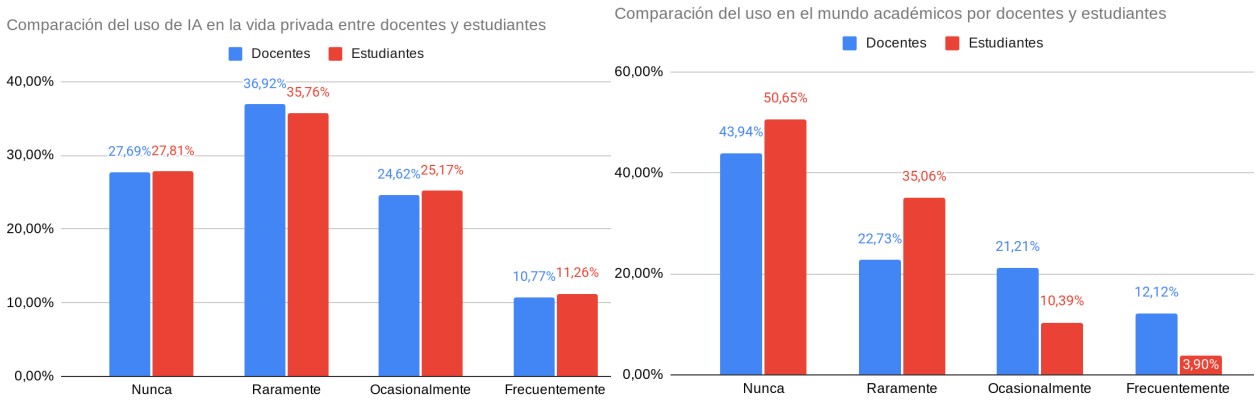


Fuente: elaboración propia.

7.1.3. Puesta en común de ambos cuestionarios. Comparación del uso y percepción de ChatGPT entre estudiante-docente.

Figura 18

Gráficas comparativas entre la frecuencia de uso de profesores y estudiantes en el ámbito personal y académico



Fuente: elaboración propia.

En esta figura (1= nunca, 5= a diario) se observa que el alumnado usa significativamente más la IA en el contexto educativo que el profesorado. Para comprobar si existen diferencias significativas entre el uso que se le da a la IA por parte del profesorado y alumnado se lleva a cabo una prueba de normalidad (tabla 20).

Tabla 20

Estudio de la prueba de normalidad aplicada al alumnado y profesorado en función del uso proporcionado a la IA

Pruebas de normalidad							
		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Categoría	Estadístico	gl	Sig.	Estadístico	gl	Sig.
USO	Profesores	,267	66	,000	,810	66	,000
	Alumnado	,288	109	,000	,786	109	,000

b. Corrección de significación de Lilliefors

Ambas categorías resultan normales pues dan un valor mayor de 0,05 en el test de normalidad. En el Anexo A se recogen los estadísticos descriptivos del USO de la IA vs Categoría (Profesorado/alumnado) cuya gráfica de barras y bigotes se discute aquí (figura 18). En ella se puede observar que no se cumple nuestra primera hipótesis (1.1)

Tabla 21

Descripción de las distintas variables estudiadas con respecto a la relación de uso de IA y el equipo docente

Descriptivos

	Categoría			Estadístico	Desv. Error
USO	Profesores	Media		2,02	,132
		95% de intervalo de confianza para la media	Límite inferior	1,75	
			Límite superior	2,28	
		Media recortada al 5%		1,96	
		Mediana		2,00	
		Varianza		1,154	
		Desv. Desviación		1,074	
		Mínimo		1	
		Máximo		4	
		Rango		3	
		Rango intercuartil		2	
		Asimetría		,584	,295
		Curtosis		-1,012	,582
	Alumnado	Media		2,74	,081
		95% de intervalo de confianza para la media	Límite inferior	2,58	
			Límite superior	2,90	
		Media recortada al 5%		2,67	
		Mediana		3,00	
		Varianza		,711	

Desv. Desviación	,843	
Mínimo	2	
Máximo	5	
Rango	3	
Rango intercuartil	1	
Asimetría	,897	,231
Curtosis	,006	,459

Figura 19

Representación gráfica del uso de la IA en relación al profesorado y alumnado

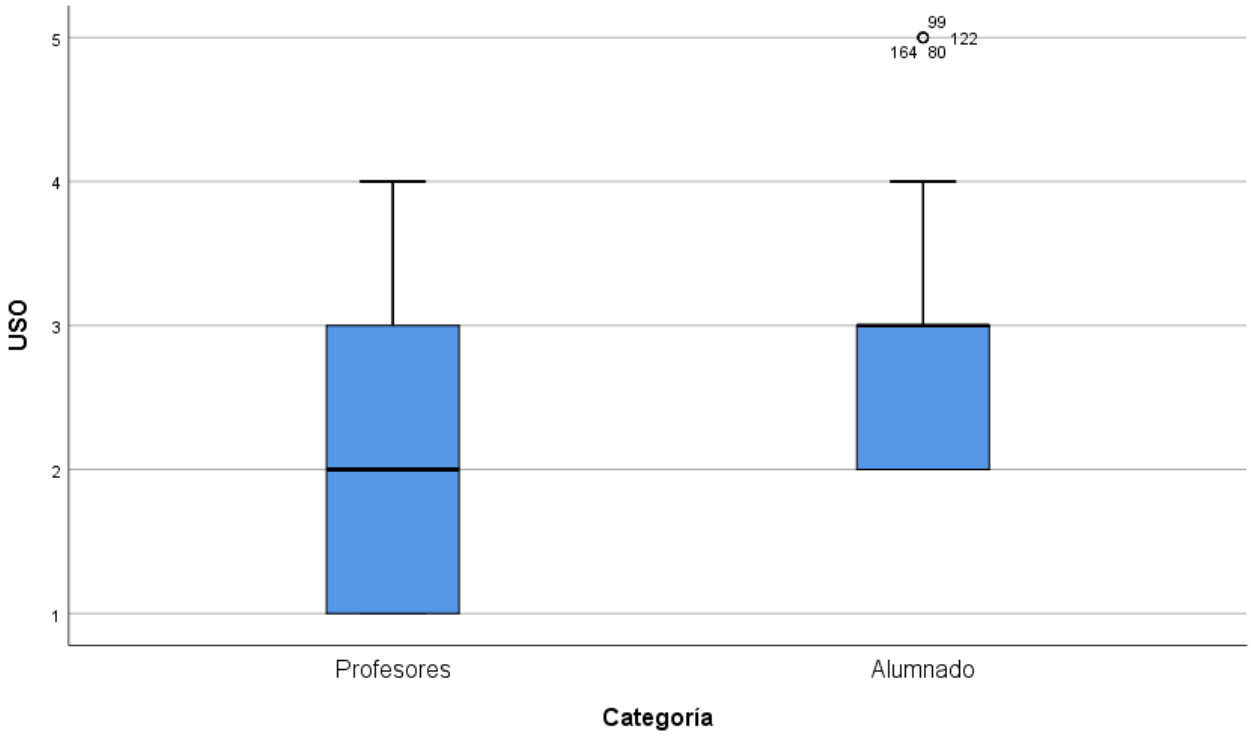


Tabla 22

Relación entre el uso de la IA y los diversos grupos encuestados

ANOVA

USO

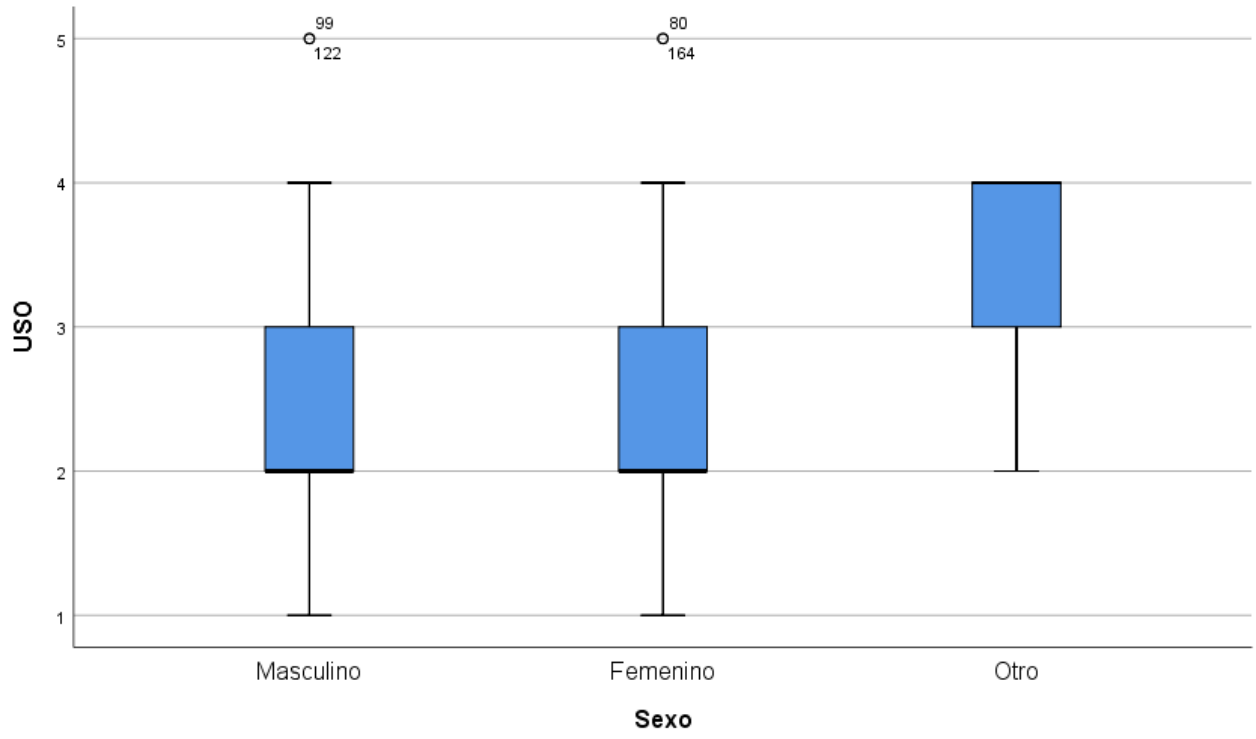
	Suma de	gl	Media	F	Sig.

	cuadrados		cuadrática		
Entre grupos	5,586	2	2,793	2,860	,060
Dentro de grupos	167,991	172	,977		
Total	173,577	174			

Al ser una significancia entre grupos mayor al 0,05 no existen diferencias significativas, (si se toma el género como una variable dicotómica) lo que cumpliría nuestra tercera hipótesis (1.3). Sin embargo tal como se observa en el gráfico de cajas y bigotes, aquellas personas que se identifican con un género no binario presentan diferencias estadísticamente significativas con un mayor uso de la IA en el contexto educativo.

Figura 20

Representación gráfica de la relación entre el uso de la IA y el sexo



7.2. Discusión sobre las respuestas de ChatGPT-3.5 a los exámenes de filosofía.

En el examen sobre el **surgimiento de la filosofía**, obtiene una nota final de 5,5. En los Anexos 2 y 3 se pueden ver las preguntas y el proceso de evaluación. En relación con sus respuestas, se puede observar cómo tiene un alto entendimiento del lenguaje humano y la comprensión lectora. Esto se evidencia a lo largo de todo el examen, siendo más notable en el primer y segundo ejercicio. En ellos se pide realizar un resumen y situar el contexto histórico, respectivamente.

En el primero, se desempeña prácticamente a la perfección, siendo el único problema la extensión, a la cual no logra adaptarse. En el segundo, se puede ver cómo sitúa correctamente el contexto histórico al que hace referencia el texto, pero falla, sin embargo, en pensar más allá de este para hablar del contexto del autor (aun cuando se da la pista de fijarse en el autor para desarrollar su respuesta). En esta pregunta se esperaba

que el modelo identificase que el texto está hablando sobre otro periodo histórico que no se corresponde al de la realidad del autor, con el objetivo de poder comentar ambos e identificar las características que pueden condicionar el pensamiento de este o de la época.

En conclusión, el modelo presenta dificultades en el razonamiento abstracto y contextual, que requiere pensar más allá del texto y realizar conexiones más complejas entre conceptos.

A partir del tercer ejercicio, y debido a esta limitación, se puede observar un empeoramiento en los resultados. La tercera pregunta está directamente ligada a la segunda, ya que, para determinar y conocer las preguntas que se plantean en alguno de los dos contextos históricos, el modelo necesita identificar ambos contextos más allá del texto. Esto le permitiría relacionarlos con un conocimiento más amplio sobre las inquietudes filosóficas que surgen en el siglo XX o durante la filosofía presocrática. Sin embargo, el modelo solo es capaz de formular preguntas relacionadas directamente con el texto.

En la cuarta pregunta se solicita situar la pregunta «¿Para qué la filosofía?» en la disciplina filosófica correspondiente. Para responder adecuadamente, el modelo debe ser capaz de distinguir entre las diferentes disciplinas existentes en la filosofía. En este caso se observa un error más significativo, ya que el razonamiento del modelo es incorrecto. Identifica esta pregunta como perteneciente a una disciplina filosófica, pero al justificar su respuesta, describe características que en realidad corresponden a otra disciplina diferente.

Finalmente, en el último ejercicio se ve nuevamente como explica correctamente todo lo que conlleva extraer información del texto y sin embargo no logra hacer una relación histórica relevante fuera de este.

En la **disertación** obtiene una nota final de 4. Para calificar esta prueba se sigue una rúbrica actual empleada para calificar a los propios alumnos de 2º de Bachiller. El modelo es calificado negativamente debido a varias deficiencias: un planteamiento del tema pobre, la ausencia de referencias a saberes básicos o al conocimiento general de otros autores, la falta de un contraargumento completo o parcial, un vocabulario pobre en términos filosóficos (teniendo en cuenta que se trata de una disertación filosófica), una estructura poco clara y, finalmente, la falta de originalidad. En la introducción, se considera que el modelo expone su tesis al mencionar la importancia del tema en función de los beneficios que aporta al ser humano. Sin embargo, no logra relacionar ni contextualizar esta idea en el marco de la actualidad, lo que hace que la introducción resulte algo pobre y sin profundidad. En cuanto a la falta de originalidad y la ausencia de conceptos o saberes básicos, el texto presenta una deficiencia importante: carece de una verdadera argumentación crítica, característica fundamental de una disertación filosófica. En lugar de desarrollar ideas propias o profundizar en el análisis, el modelo se limita a parafrasear las ideas del pequeño texto proporcionado en el input, usando las reflexiones de Russell. El modelo tiene una dependencia excesiva en el recurso y una falta de elaboración personal.

7.3. Discusión sobre las respuestas de ChatGPT-4 a los exámenes de filosofía.

En el examen sobre el surgimiento de la filosofía obtiene una nota final de 9,4. El primer ejercicio que requiere de una comprensión textual lo realiza correctamente. Se le penaliza ligeramente en el apartado de vocabulario debido a que no menciona explícitamente los términos «mito» y «logos». Aunque estos conceptos no aparecen de manera literal en el texto, representan el proceso que este describe. Sin embargo, esta omisión no se considera un error propiamente dicho, sino más bien una falta de precisión, valorada bajo un criterio exigente.

En el segundo ejercicio identifica adecuadamente tanto el contexto del texto como el del autor (al especificar que se fije en el autor en un segundo input). En el tercer ejercicio piensa adecuadamente en los problemas del contexto histórico y genera correctamente preguntas en la disciplina filosófica. En la cuarta pregunta justifica a la perfección su respuesta. En la quinta define correctamente y lo relaciona con los acontecimientos históricos.
En conclusión, presenta una buena comprensión del lenguaje humano y tiene la capacidad de relacionar conceptos fuera del texto inicial y razonar.

En la disertación obtiene una nota final de 5,5, siendo penalizado por un planteamiento del tema pobre, la ausencia de referencias a saberes básicos o al conocimiento general de otros autores, la falta de un contraargumento completo o parcial, un vocabulario pobre en términos filosóficos y la falta de originalidad. Al igual que el modelo anterior, en la introducción, se considera que expone su tesis al mencionar la importancia del tema en función de los beneficios que aporta al ser humano. Sin embargo, no logra relacionar ni contextualizar esta idea en el marco de la actualidad, lo que hace que la introducción resulte algo pobre y sin profundidad. En cuanto a la falta de originalidad y la ausencia de conceptos o saberes básicos, el texto carece de una verdadera argumentación crítica. En lugar de desarrollar ideas propias o

profundizar en el análisis, se limita a parafrasear las ideas del texto proporcionado en el input, usando las reflexiones de Russell. No hay elaboración de ideas propias. No obstante, la estructura es clara.

Exámenes de EVAU:

En el examen EVAU Madrid 2023 Julio Opción A obtiene una nota final de 5,1.

En el examen EVAU Madrid 2023 Julio Opción B obtiene una nota final de 6,6

En el examen EVAU Madrid 2021 Junio Opción A obtiene una nota final de 7,8

En el examen EVAU Madrid 2021 Junio Opción B obtiene una nota final de 4,75

En el examen EVAU Madrid 2021 Julio Opción A obtiene una nota final de 6,5

En el examen EVAU Madrid 2021 Julio Opción B obtiene una nota final de 6,1

- Ejercicio tipo 1 de EVAU:

En este tipo de ejercicio se pide exponer las ideas fundamentales del texto propuesto y la relación que existe entre ellas. Primero, se le penaliza por no detallar en las características o el contexto del autor al que pertenece el texto (EVAU Madrid 2023 Julio Opción A, EVAU Madrid 2021 Junio Opción A). En segundo lugar, se le penaliza por no mencionar ideas fundamentales del texto según los criterios de corrección de la propia EVAU o por no desarrollarlas con la adecuada profundidad. Por ejemplo, en el examen de EVAU Madrid 2023 Julio Opción B no menciona que existen otras vías en la filosofía de Tomás de Aquino y solo explica la que aparece en el texto. Tampoco menciona la estructura argumental de las vías ni sus tres partes fundamentales (datos del mundo sensible, desarrollo argumental, y conclusión). En la EVAU Madrid 2021 Junio Opción A, menciona el escepticismo metódico pero no nombra las cuatro fases. En la EVAU Madrid 2021 Junio Opción B, omite la mayoría de ideas fundamentales en la filosofía de Marx, centrándose solo en el lenguaje y la conciencia como producto social, así, omite ideas como el rechazo a la división cronológica de la historia propia del filósofo, su materialismo histórico en cuatro fases (tribal (parentesco), antigua (estatal), feudal y capitalista) o la idea de comunismo. En la EVAU Madrid 2021 Julio Opción A, no menciona el Republicanismo, entre otras. También sufre penalizaciones en cuanto a vocabulario puesto que a veces no menciona términos esenciales en la filosofía del autor, como son los nombres de fases o conceptos.

- Ejercicio tipo 2 (2, 3, 4 de los exámenes) de EVAU:

En este tipo de ejercicio se pide exponer un problema concreto sobre un autor de una época concreta. Se le penaliza al olvidar mencionar conceptos fundamentales, al no emplear un vocabulario o mencionar términos esenciales en la filosofía del autor o al elegir uno que no pertenece a la época.

Respecto a carencias en el contenido, en la EVAU Madrid 2023 Julio Opción A ejercicio 2 habla del «fin último» y las virtudes de Tomás de Aquino, pero no distingue claramente entre la vía teórica y la vía del hábito. En la EVAU Madrid 2023 Julio Opción B ejercicio 2, no explica los niveles en profundidad del dualismo epistemológico de Platón (doxa (pístis, eikasia) y episteme (diánoia, nóesis)). En el ejercicio 2 de la EVAU Madrid 2021 Junio Opción A menciona las virtudes aunque no los términos secundarios ni el alcance de la excelencia final en la filosofía de Aristóteles. En el ejercicio 3 de ese mismo examen, no menciona el modelo de gobierno republicano en la filosofía de Rousseau. También, menciona ideas que realmente no eran necesarias en este apartado como la política o la razón como naturaleza del hombre, sin embargo, esto es correcto aunque no se evalúe.

En la EVAU Madrid 2021 Junio Opción B en el ejercicio 2, el modelo no trata el problema del ser humano sino que habla solamente del intento de Santo Tomás de Aquino en conciliar fé y razón. En el ejercicio 3, no menciona el giro copernicano en la epistemología con Kant. En el 4, no menciona las fases de la evolución hacia el superhombre de Nietzsche.

En el examen EVAU Madrid 2021 Julio Opción B en el ejercicio 2, no menciona directamente la virtud para alcanzar la excelencia (Aristóteles) ni la justicia según el filósofo o los tipos de virtudes (intelectuales y morales).

En algunos exámenes escoge hablar de otros filósofos, en el caso de la EVAU Madrid 2023 Julio Opción A ejercicio 3 escoge hablar sobre Voltaire que no es el autor más representativo aunque desarrolla bien algunas ideas. En la EVAU Madrid 2023 Julio Opción B ejercicio 3 escoge hablar del contrato social de Hobbes y lo hace bien, aunque la relación entre el contexto histórico y su teoría no está completamente desarrollada. Asimismo, en el ejercicio 4 escoge a Peter Singer, que es de la época contemporánea que no tan relevante en la ética por un enfoque más pragmático.

En la EVAU Madrid 2021 Junio Opción A ejercicio 4, escoge a Sartre en vez de a Nietzsche. Esta elección es adecuada y se desarrolla bien, aunque hubiera estado bien que se enumerará en el concepto de «condición humana» de Sartre los límites comunes a todos los hombres, que son: 1. estar arrojado en el mundo; 2. tener que trabajar; 3. vivir en medio de los demás; 4. ser mortal. O al menos que se destacara el carácter inevitable. El resto de conceptos como la angustia o la mala fe se tratan satisfactoriamente.

En la EVAU Madrid 2021 Julio Opción A ejercicio 4, escoge hablar de Kant de forma incorrecta. Kant pertenece a la filosofía moderna y no a la contemporánea Aunque el modelo lo sabe decide hablar de él.

En la EVAU Madrid 2021 Julio Opción B ejercicio 3, escoge hablar de Sartre que es contemporáneo, no moderno.

En conclusión, menciona bastantes de las ideas fundamentales de cada autor aunque se deja algunas de vez en cuando. Los fallos más graves los comete al seleccionar un autor que no es de la época que se pide. Esto le ocurre al modelo en dos exámenes y puede considerarse el error más grave ya que un alumno que no conoce profundamente sobre los filósofos, no lee atentamente e interpreta la respuesta o no verifica su veracidad puede no identificar este error.

7.4. Discusión sobre las respuestas de ChatGPT-4o a los exámenes de filosofía.

En el examen sobre el surgimiento de la filosofía obtiene una nota final de 9,2. En el primer ejercicio realiza bien el resumen. En el segundo, identifica adecuadamente el contexto del texto y sabe identificar a otros filósofos importantes de la época contemporánea. Además, lo relaciona incluso con el contexto histórico en España. Sin embargo, aunque en términos temporales es correcto, siendo una evaluación de filosofía, se espera una mayor claridad o mención explícita del periodo contemporáneo. El cuarto y quinto ejercicio los responde también satisfactoriamente. En resumen, el modelo muestra un sólido entendimiento del lenguaje humano y es capaz de conectar ideas externas al texto inicial, razonando de forma adecuada.

En la disertación obtiene una nota final de 5,75. Se le resta nota debido a un planteamiento del tema pobre, la ausencia de referencias a saberes básicos o al conocimiento general de otros autores, un vocabulario pobre en términos filosóficos, la falta de originalidad y la estructura. Al igual que el modelo anterior, en la introducción, se considera que expone su tesis al mencionar la importancia del tema en función de los beneficios que aporta al ser humano. Sin embargo, no logra relacionar ni contextualizar esta idea en el marco de la actualidad, lo que hace que la introducción resulte algo pobre y sin profundidad. En cuanto a la falta de originalidad y la ausencia de conceptos o saberes básicos, el texto carece de una verdadera argumentación crítica. En lugar de desarrollar ideas propias o profundizar en el análisis, se limita a parafrasear las ideas del texto proporcionado en el input, usando las reflexiones de Russell. No hay elaboración de ideas propias.

Por otro lado, este modelo introduce lo que podría considerarse un contraargumento parcial «...el ocio se asocia a menudo con el consumo pasivo» y emplea un vocabulario algo más rico «alienación, paradoja…».

En el examen EVAU Madrid 2023 Julio Opción A obtiene una nota final de 8,15
En el examen EVAU Madrid 2023 Julio Opción B obtiene una nota final de 6,3
En el examen EVAU Madrid 2021 Junio Opción A obtiene una nota final de 8,7
En el examen EVAU Madrid 2021 Junio Opción B obtiene una nota final de 5,7
En el examen EVAU Madrid 2021 Julio Opción A obtiene una nota final de 7,4
En el examen EVAU Madrid 2021 Julio Opción B obtiene una nota final de 9,1

- Ejercicio tipo 1 de EVAU:

En este tipo de ejercicio se pide exponer las ideas fundamentales del texto propuesto y la relación que existe entre ellas. Primero, se le penaliza en los exámenes por no explicar brevemente el contexto del autor o quien es en una adecuada introducción (EVAU Madrid 2023 Julio Opción A, EVAU Madrid 2021 Junio opción A). En segundo lugar, se le penaliza por no mencionar ideas fundamentales del texto según los criterios de corrección de la propia EVAU o por no desarrollarlas con la adecuada profundidad. Por ejemplo, en el examen de EVAU Madrid 2023 Julio Opción B no incluye una enumeración de las cinco vías ni explica sus funciones y falta profundidad en la explicación de la ejemplificación del proceso de argumentación de la cuarta vía en el texto. Tampoco menciona las tres partes de las vías aunque sí la estructura y el carácter.

En la EVAU Madrid 2021 Junio Opción B, sobre Descartes, sobre Marx, omite casi todas las idea y solo explica el lenguaje y la conciencia como producto social.

- Ejercicio tipo 2 (2, 3, 4 de los exámenes) de EVAU:

En este tipo de ejercicio se pide exponer un problema concreto sobre un autor de una época concreta. Se le penaliza al olvidar mencionar conceptos fundamentales, al no emplear un vocabulario o mencionar términos esenciales en la filosofía del autor o al elegir uno que no pertenece a la época.

Respecto a carencias en el contenido, en la EVAU Madrid 2021 Junio Opción A ejercicio 2 a modo de experimento al ver la diferencia de extensión entre este y el modelo anterior se le preguntó primero sin límite de extensión y luego con límite de 210 palabras para observar si mantiene todas las ideas. Sin el límite mencionó todo a la perfección. Con el límite no diferenció explícitamente entre la vía teórica (contemplación beatífica de Dios como felicidad perfecta) y la vía del hábito (virtudes morales y acción virtuosa).

En la EVAU Madrid 2021 Junio Opción B ejercicio 2, se deja dos ideas esenciales aunque habla incluso de las tres facultades del alma que no eran necesarias (filosofía de Santo Tomás de Aquino). En el ejercicio 3 no menciona el giro copernicano en la epistemología de Kant tampoco menciona la crítica a los imperativos hipotéticos. En este también profundiza positivamente de más en la formulación del imperativo categórico mencionando incluso algunas de las formulaciones textualmente. En el ejercicio 4, no menciona las fases de la evolución hacia el superhombre de Nietzsche.

En la EVAU Madrid 2021 Julio Opción A ejercicio 2, solo menciona el alma racional en Tomás de Aquino ni resalta su importancia en la esencia del humano.

En el examen EVAU Madrid 2021 Julio Opción B en el ejercicio 2, no menciona directamente la virtud para alcanzar la excelencia (Aristóteles) según el filósofo o los tipos de virtudes (intelectuales y morales).

En algunos exámenes escoge hablar de otros filósofos, en la EVAU Madrid 2023 Julio Opción B ejercicio 3, escoge hablar del contrato social de Hobbes de forma correcta. En la EVAU Madrid 2023 Julio Opción B ejercicio 4, escoge a Kant que no sería una elección adecuada ya que pertenece a la época moderna y no a la contemporánea. En la EVAU Madrid 2021 Junio Opción A ejercicio 3, escoge hablar del contrato social de Hobbes que es previo al de Rousseau, pero es correcto. En el ejercicio 4 del mismo examen, el modelo escoge hablar de Sartre también correctamente.

En la EVAU Madrid 2021 Julio Opción A ejercicio 4, escoge hablar de Kant teniendo en cuenta que es moderno y considerando que influye en la filosofía contemporánea. Sin embargo, se pide que sea de la época contemporánea exactamente y está mal.

En el examen EVAU Madrid 2021 Julio Opción B, escoge hablar de Descartes, que también es un filósofo moderno. Lo explica correctamente. En el 4, escoge hablar de Habermas, filósofo también contemporáneo. Lo desarrolla bien.

En conclusión, menciona bastantes de las ideas fundamentales de cada autor aunque se deja algunas de vez en cuando. Los fallos más graves cometido son en la EVAU Madrid 2021 Junio Opción A ejercicio 2, en el que menciona los tipos de virtudes de Aristóteles y pone ejemplos. Así pues, explica que ambas se adquieren con la práctica (en realidad una se adquiere con la educación siendo más preciso). Esto podría considerarse un pequeño error de concepto. También, en uno de los exámenes habla de un autor que no pertenece a la época que se le pregunta. Por otro lado, su desarrollo de ideas es muy claro. Asimismo, el primer modelo explica erróneamente un autor que no pertenece a la época que se pide con mayor frecuencia.

7.5. Discusión sobre las respuestas de ChatGPT-3.5 a los problemas de 3º ESO.

En la resolución del problema de física se puede observar como ChatGPT desglosa todos los pasos para llegar a la solución del problema. Sin embargo, se le pide calcular la aceleración final, la distancia recorrida y el aumento del desplazamiento y este confunde los datos del enunciado, lo que le induce a error tanto en la resolución como en la gráfica plateada. Una vez alertada el error, esta IA intenta corregirlo pero vuelve a confundir el término de «desplazamiento» y la «distancia recorrida» (términos que confunden usualmente los alumnos). Con tan solo estos errores mencionados anteriormente, se analiza la falta de precisión, claridad y comprensión que presenta esta herramienta, en este caso, debido a que comete errores muy usuales y hay que indicarle el fallo para que comprenda qué es lo que está haciendo mal.

Asimismo, presenta un pobre análisis en el *prompt* dado, ya que no tiene en cuenta toda la información aportada (en este caso del enunciado). Esto se observa cuando en el primer intento de la resolución del problema, no tiene en cuenta que la velocidad es constante. En el tercer intento de la resolución del *inverse problem*, es alarmante el hecho de que no tiene en cuenta las leyes de la física y sigue confundiendo términos. Por ello, hay que señalarle que cuando un objeto se mueve en dirección contraria también cubre una distancia. Este es otro claro ejemplo de la falta de comprensión y la carencia de adaptabilidad que tiene esta herramienta frente al usuario.

Nuevamente, vuelve a cometer fallos de términos (con el desplazamiento) y cambios constantes en los resultados de la posición final.

En el siguiente apartado, se le pide a ChatGPT una gráfica de velocidad frente a tiempo y vuelve a cometer un fallo muy grave a la hora de trazar la gráfica. Debido a que la partícula en movimiento desacelera, la gráfica debe de ir hacia abajo, pero este la traza hacia arriba. El usuario le señala el fallo, sin embargo, vuelve a trazar mal la gráfica. Posteriormente, vuelve a cometer varios errores en la realización de la gráfica.

En conclusión, ChatGPT no tiene muy claros los conceptos físicos y carece de un análisis total en la información dada. Asimismo, no toma muy bien las correcciones y comete muchos fallos, indicando la carencia de precisión y habilidad para gestionar los errores.

8. Discusión sobre las respuestas de las versiones ChatGPT-4 a los problemas de 3º ESO.

En este apartado se muestra la discusión sobre las respuestas que proporcionan los modelos ChatGPT-4, ChatGPT-4.o y ChatGPT-o1 preview. En el caso de las discusiones de las respuestas de ChatGPT-4.o se compararán las respuestas que proporciona tanto en inglés como en español además de comparar los resultados que se obtienen desde la página web oficial de OpenAI y desde el buscador alternativo you.com.

8.1. Discusión sobre las respuestas de ChatGPT-4o a los problemas de 3º ESO.

Al someter a ChatGPT-4.o a una serie de ejercicios de Física de 3ºESO vía *you.com* y la página oficial de *OpenAI* en inglés y en español, se examinan sus correspondientes respuestas:

En el primer ejercicio se le pide a la IA dibujar una gráfica velocidad-tiempo y encontrar la aceleración y la distancia recorrida por el automóvil en el tiempo de 10 s. Se puede observar que la herramienta disponible en OpenAI llega a la resolución correcta de todos los apartados sin cometer ningún fallo. Asimismo, la representación gráfica del ejercicio es correcta y los resultados dados por este mismo modelo, pero con el enunciado en español no varían. En el caso de la herramienta accesible en you.com, ChatGPT-4.o proporciona los mismos correctos valores numéricos tanto en inglés como en español. Aparentemente, tanto la gráfica realizada en español como la gráfica realizada en inglés son correctas, aunque en términos de claridad, la representación gráfica del ejercicio en español tiende a confusión.

En el segundo ejercicio, la herramienta disponible en OpenAI debe trazar una gráfica velocidad-tiempo y hallar la distancia recorrida por el tren mientras desacelera. Se observa que es capaz de solucionar el problema sin ningún fallo y traza correctamente la gráfica velocidad-tiempo. En el caso de you.com muestra la misma capacidad de resolución ante este mismo ejercicio, es decir, llega correctamente a los valores de los resultados de cada apartado. En cuanto a la resolución del problema en español, ambos modelos llegan al resultado numérico adecuado. No obstante, ambas gráficas realizadas por you.com tanto en inglés como en español están incorrectamente representadas.

En el tercer ejercicio, se le pide a la herramienta disponible en *OpenAI* trazar una gráfica velocidad-tiempo y encontrar la distancia total recorrida por la partícula. Nuevamente, halla el resultado del enunciado sin ningún fallo y traza correctamente la gráfica velocidad-tiempo. Lo mismo ocurre en la resolución del problema en la herramienta disponible en you.con. En cuanto a la resolución del problema numéricamente en español, ambos modelos llegan al correcto resultado. No obstante, llama la atención que aunque los valores numéricos hayan sido calculados correctamente, las gráficas que muestran you.com tanto en español como en inglés son erróneas y bastante confusas. En cambio, la IA accesible en OpenAI representa bastante bien la gráfica.

En el cuarto ejercicio, ChatGPT-4.o disponible en OpenAI debe representar una gráfica velocidad-tiempo y hallar la aceleración final y el desplazamiento final de la partícula desde su punto inicial. Esta herramienta llega a los resultados correctos sin necesidad de introducir *prompts* adicionales y a una adecuada representación gráfica del ejercicio. El modelo accesible en you.com muestra la misma capacidad de resolución, sin embargo, no realiza una correcta representación gráfica del problema. En cuanto a la resolución del problema en español, ambos modelos llegan al correcto resultado numérico. No obstante, GPT-4o tampoco realiza una correcta representación gráfica en español.

En el quinto ejercicio, la herramienta de OpenAI debe llegar correctamente a la velocidad final, la distancia recorrida y el incremento del desplazamiento. Asimismo, debe trazar una gráfica velocidad-tiempo para el intervalo de 10 segundos. Este modelo llega correctamente a todos los valores pedidos en los distintos apartados sin la necesidad de introducir nuevos *prompts* y traza correctamente la gráfica pedida. En el caso de you.com se observa que la herramienta tiene más problemas a la hora de resolver los ejercicios, aunque hace una representación gráfica correcta del ejercicio. Este requiere un *prompt* adicional para llegar al resultado final del valor de la distancia recorrida pedido. Sin embargo, se observa que es capaz de corregir su error y llegar al resultado correcto. Cabe destacar que el error cometido por la herramienta fue por un descuido en el signo del proceso de cálculo. En cambio, el modelo You.com (al igual que el modelo OpenAI) llega al resultado correcto al primer intento cuando se le pregunta en español sin necesidad de *prompts* adicionales y la representación gráfica es adecuada.

En el sexto ejercicio, la herramienta debe calcular la aceleración final, la distancia recorrida, el aumento del desplazamiento y trazar una gráfica velocidad-tiempo y encuentre, para el segundo intervalo. Se observa que la IA disponible en OpenAI tiene bastantes complicaciones a la hora de llegar al resultado correcto. En

primer lugar, falla al calcular la aceleración final ya que esta no tiene en cuenta que la aceleración final del segundo intervalo es la aceleración inicial del tercer intervalo, por consiguiente necesita tres *prompts* adicionales para clarificar este apartado. Además, realiza un cálculo erróneo que ya se habían indicado en los *prompts* anteriores. No obstante, esta no ubica su error por lo que se introducen tres *prompts* más, indicando cual debería de ser el resultado adecuado para ver si es capaz de observar su error. Debido a todos estos errores de cálculo, la representación gráfica pedida tampoco es correcta. En conclusión, ChatGPT-4.o en OpenAI necesita un total de siete *prompts* para llegar a la respuesta correcta. Asimismo, la herramienta accesible en you.com tampoco es capaz de llegar a los resultados correctos. Después de 5 *prompts* adicionales indicando el error del apartado y proporcionando los resultados correctos, es incapaz de llegar a la respuesta correcta de la distancia recorrida. Por ende, la representación gráfica también es incorrecta. En cuanto a las respuestas proporcionadas en español, no difieren con las respuestas generadas en inglés, es decir, necesitan *prompts* adicionales y los resultados finales son incorrectos.

En el séptimo ejercicio se le pide a la IA que resuelva a partir de una gráfica que debe elaborar él: la distancia recorrida por un avión a los 5 segundos, a los T segundos, el gradiente de la línea y la aceleración del avión. En este ejercicio, Chat GPT-4.o disponible en you.com realiza una resolución completamente correcta de los tres apartados sin necesidad de brindarle *prompts* extra. La IA oficial de OpenAI también llega a la resolución correcta del ejercicio. Además, cabe destacar que en este ejercicio ambos modelos llegan a trazar la misma gráfica y de manera correcta. En cuanto a la resolución del problema en español, ambos modelos llegan al correcto resultado.

En el octavo ejercicio se le pide a ChatGPT-4.o trazar una gráfica de velocidad frente al tiempo de la pelota, buscar una fórmula para la distancia recorrida por la pelota en *t* segundos y el tiempo que tarda la pelota en llegar al suelo. Nuevamente, ambos modelos accesibles en los destinos programas realizan una resolución correcta de todos los apartados, sin errores en sus explicaciones y sin necesidad de proporcionar más información al respecto. No obstante trazan gráficas diferentes. La gráfica que se muestra en you.com es errónea y además no presenta mucha claridad como ya se ha mencionado anteriormente en los demás ejercicios. En el caso del ejercicio realizado en español, el modelo de ambos programas responden correctamente a todos los apartados del ejercicio y las gráficas también son correctas.

En el último ejercicio se le pide a la herramienta que trace una gráfica que muestre el incremento de la velocidad suponiendo que el tiempo en el que el coche empieza a acelerar es de 0 segundos, encontrar la aceleración del coche cuando t = 0, encontrar la distancia viajada en dos situaciones temporales distintas y trazar una gráfica de la distancia frente al tiempo del coche. Nuevamente, ambos programas realizan una resolución numérica del ejercicio correcta tanto en español como en inglés. La representación gráfica del modelo accesible en la página oficial también es correcta. No obstante, las gráficas trazadas en *you.com* son erróneas.

8.2. Discusión sobre las respuestas de ChatGPT-o1 preview a los problemas de 3º ESO.

En el primer ejercicio se le pide a la IA dibujar una gráfica velocidad-tiempo y encontrar la aceleración y la distancia recorrida por el automóvil en el tiempo de 10 s. Se puede observar que esta llega a la resolución correcta de todos los apartados sin cometer ningún fallo y la descripción gráfica que realiza es adecuada.

En el segundo ejercicio se le pide a la herramienta trazar una gráfica velocidad-tiempo y hallar la distancia recorrida por el tren mientras desacelera. Nuevamente, la herramienta llega a los resultados correctos sin necesidad de introducir *prompts* adicionales y a una buena descripción gráfica.

En el tercer ejercicio, se le pide a la IA trazar una gráfica velocidad-tiempo y buscar la distancia total recorrida por la partícula. En este ejercicio tampoco presenta dificultades en llegar al resultado esperado y realiza una correcta descripción gráfica.

En el cuarto ejercicio, debe calcular representar una gráfica velocidad-tiempo y hallar la aceleración final y el desplazamiento final de la partícula desde su punto inicial. En este ejercicio realiza una correcta resolución del problema y una correcta descripción gráfica.

En el quinto ejercicio, se le pide a la herramienta calcular la velocidad final, la distancia recorrida y el incremento del desplazamiento. Asimismo, debe trazar una gráfica velocidad-tiempo para el intervalo de 10 segundos. No obstante, se observa que ChatGPT-o1 preview tiene varias complicaciones y confusión a la hora de calcular la distancia recorrida por la partícula. A pesar de introducir dos *prompts* adicionales indicando que ha realizado mal los cálculos del apartado, no consigue ver dónde comete el error y acaba dando el mismo resultado que erróneo que daba al principio. En el cuarto *promt* se le indica cuál es el resultado que debe obtener y en ese instante logra comprender su error y añade un párrafo adicional en el que explica que cometió el error en el cálculo de la distancia recorrida en el tender intervalo. Sin embargo,

cambia el resultado del tercer apartado (cálculo del incremento del desplazamiento) que inicialmente estaba correcto. Por consiguiente, se le vuelve a introducir un quinto *prompt* indicando el nuevo fallo en sus cálculos y proporcionando el resultado correcto, pero al corregir este nuevo error de cálculo vuelve a dar mal el resultado del segundo apartado (cálculo de la distancia recorrida). No llega a comprender el error que está cometiendo y por lo tanto no llega al resultado correcto.

En el sexto ejercicio, se le pide la ChatGPT-o1 preview calcular la aceleración, la distancia recorrida del problema, el incremento del desplazamiento y trazar una gráfica velocidad-tiempo. En la resolución de este ejercicio también presenta algunas complicaciones que le impiden llegar a los resultados correctos. En el primer intento calcula correctamente la aceleración final, pero da resultados erróneos en el segundo y tercer apartado. Por ello, se le introduce un nuevo *prompt* indicando que ha calculado mal los correspondientes apartados. Sin embargo, vuelve a calcular todos los apartados mal, incluido el primero (que ya lo tenía bien). Por consiguiente, se le vuelve a introducir un nuevo *promt*, pero no logra llegar a los resultados correctos. Esta vez se le introduce otro *prompt* indicando el valor correcto de la aceleración para comprobar si a partir de ese dato consigue llegar a todos los resultados correctos de los demás apartados, pero no lo logra. Así que se le introduce un quinto *prompt* indicando el resultado correcto de los dos últimos apartados con el fin de que vea su error en el cálculo. Llega al resultado correcto, pero porque afirma que ese resultado corresponde a la distancia total y el incremento del desplazamiento del movimiento total de toda la partícula. No obstante, esa afirmación es incorrecta por lo que el resultado tampoco es correcto, ya que no está contestando a lo pedido (la distancia recorrida y el incremento del desplazamiento durante el segundo intervalo).

En el séptimo ejercicio, la ChatGPT-o1 preview realiza el ejercicio bastante bien. Se le pide que calcule a partir de una gráfica que debe elaborar él: la distancia recorrida por un avión a los 5 segundos, a los T segundos, el gradiente de la línea y la aceleración del avión. En este ejercicio no muestra ninguna complicación y resuelve todos los apartados sin ningún fallo y realiza una correcta descripción gráfica..

En el octavo ejercicio, se le pide a la herramienta trazar una gráfica de velocidad frente al tiempo de la pelota, buscar una fórmula para la distancia recorrida por la pelota en *t* segundos y el tiempo que tarda la pelota en llegar al suelo. En este caso, resuelve el ejercicio perfectamente.

En el último ejercicio, se le pide a la herramienta que trace una gráfica que muestre el incremento de la velocidad suponiendo que el tiempo en el que el coche empieza a acelerar es de 0 segundos, encontrar la aceleración del coche cuando t = 0, encontrar la distancia viajada en dos situaciones temporales distintas y trazar una gráfica de la distancia frente al tiempo del coche. Nuevamente, vuelve a realizar una resolución del ejercicio correcta.

8.3. Discusión sobre las respuestas de ChatGPT-4 a los problemas de 3º ESO.

En el primer ejercicio se le pide a la IA hallar dibujar una gráfica velocidad-tiempo y encontrar la aceleración y la distancia recorrida por el automóvil en el tiempo de 10 s. Se puede observar que esta llega a la resolución correcta de todos los apartados sin cometer ningún fallo. Se puede observar que esta llega a la resolución correcta de todos los apartados sin cometer ningún fallo.

En el segundo ejercicio se le pide a la herramienta trazar una gráfica velocidad-tiempo y hallar la distancia recorrida por el tren mientras desacelera. Nuevamente, ChatGPT-4 llega a los resultados correctos sin necesidad de introducir prompts adicionales.

En el tercer ejercicio, se le pide a la IA trazar una gráfica velocidad-tiempo y buscar la distancia total recorrida por la partícula. En este ejercicio tampoco presenta dificultades en llegar al resultado esperado.

En el cuarto ejercicio, debe calcular representar una gráfica velocidad-tiempo y hallar la aceleración final y el desplazamiento final de la partícula desde su punto inicial. En este ejercicio realiza una correcta resolución del problema.

En el quinto ejercicio, se le pide a la herramienta calcular la velocidad final, la distancia recorrida y el incremento del desplazamiento. Asimismo, debe trazar una gráfica velocidad-tiempo para el intervalo de 10 segundos. En este caso, llega al resultado correcto de todos los apartados.

En el sexto ejercicio la IA debe trazar una gráfica velocidad-tiempo y encontrar, para el segundo intervalo: la aceleración final, la distancia recorrida y el aumento del desplazamiento. En este ejercicio, la herramienta no llega a calcular ningún apartado correctamente por lo que se le introducen dos prompts adicionales indicando que no ha realizado bien los cálculos y tras volver a intentarlo dos veces más, no logra resolver el problema. Como consecuencia, se le introduce un tercer promt adicional indicando cuales deben

de ser los resultados correctos para analizar si la herramienta es capaz de autocorregirse y detectar su error, pero no llega a comprender cuales son sus errores. Por ello, no llega a resolver correctamente el ejercicio.

En el séptimo ejercicio se le pide a la IA que resuelva a partir de una gráfica que debe elaborar él: la distancia recorrida por un avión a los 5 segundos, a los T segundos, el gradiente de la línea y la aceleración del avión. En este ejercicio, Chat GPT-4 realiza una resolución completamente correcta de los tres apartados sin necesidad de brindarle prompts extra.

En el octavo ejercicio, se le pide a la herramienta trazar una gráfica de velocidad frente al tiempo de la pelota, buscar una fórmula para la distancia recorrida por la pelota en t segundos y el tiempo que tarda la pelota en llegar al suelo. En este ejercicio resuelve el ejercicio perfectamente.

En el último ejercicio se le pide a la herramienta que trace una gráfica que muestre el incremento de la velocidad suponiendo que el tiempo en el que el coche empieza a acelerar es de 0 segundos, encontrar la aceleración del coche cuando t = 0, encontrar la distancia viajada en dos situaciones temporales distintas y trazar una gráfica de la distancia frente al tiempo del coche. Nuevamente, vuelve a realizar una resolución del ejercicio correcta.

9. Discusión sobre las respuestas de las versiones de ChatGPT-4 a los problemas de 2º Bachiller.

En este apartado se muestra la discusión sobre las respuestas que proporcionan los modelos ChatGPT-4 y ChatGPT-4o ante los problemas extraídos de la unidad de interacción gravitatoria incluída en el temario de física en selectividad.

9.1. Discusión sobre las respuestas de ChatGPT-4 a los problemas de 2º Bachiller.

En la primera prueba se le pide al modelo realizar un dibujo de la situación con todos los detalles y calcula el vector campo gravitatorio en el punto C(4,3) m, calcular el potencial gravitatorio en los puntos C(4,3) m y D(2,2) m y calcular el trabajo necesario para llevar una masa, m_3, de 5 kg desde el punto C hasta el D y razona sobre su signo respecto a dos masas iguales, m_1 y m_2, de 100 kg están situadas en los puntos A(0,0) m y B(0,3) m. En este ejercicio se puede observar que realiza sin ninguna dificultad el primer y el segundo apartado. No obstante, en el tercer apartado se equivoca en el signo del resultado debido a un fallo de signo en el planteamiento de la ecuación del trabajo. Por ello, se le introduce un *prompt* adicional indicando su error. De este modo, la IA llega a la solución correcta del ejercicio. Además, es capaz de razonar el significado del signo.

En el segundo ejercicio, la IA debe calcular la velocidad orbital, el periodo de su órbita, la energía mecánica en la órbita y el trabajo mínimo que deberían realizar los motores del satélite si fuese necesario pasar a otra órbita más alejada, el doble de la primera: 2h con respecto a un satélite de 1350 kg destinado a la observación climática y oceanográfica que orbita La Tierra a una altura sobre su superficie de 660 km. En esta prueba se observa que la herramienta no tiene dificultades para llegar a los resultados correctos de los tres primeros apartados. Sin embargo, realiza una interpretación incorrecta de la ecuación del trabajo en el último apartado. Por ello, se le introduce un *prompt* adicional para proporcionarle la ecuación adecuada. En ese momento, llega a la resolución correcta del ejercicio.

En el tercer ejercicio, la herramienta debe calcular el periodo de revolución del satélite expresado en horas, la energía potencial y cinética del satélite en su órbita con respecto a un satélite que tiene una masa m = 500 kg y su órbita, supuesta circular, se encuentra a una distancia de $2{,}3 \cdot 10^4$ km de la superficie terrestre. En esta prueba se observa que la herramienta no tiene complicaciones en resolver el primer apartado. No obstante, al realizar el cálculo de las energías perdidas, realiza un correcto planteamiento, pero una resolución matemática errónea. Por ello, se le introduce un nuevo *prompt* adicional indicando el fallo. A pesar de esto, no logra identificar el fallo, por lo que se le introduce otro *prompt* adicional en el que se le especifica el valor del resultado correcto. En ese instante es capaz de identificar su error y corregirlo.

En el cuarto ejercicio, el modelo debe hallar la velocidad de escape desde la superficie y a qué altura, desde la superficie, alcanzará un objeto si se lanza con la mitad de la velocidad de escape, sabiendo que Marte (supuesto esférico) tiene un radio de $R = 3{,}39 \cdot 10^6$ m, y el valor de la gravedad en su superficie es g = 3,71 m/s^2. En esta prueba, la herramienta no muestra dificultad alguna para llegar a los resultados correctos del ejercicio.

En el quinto ejercicio, la herramienta debe calcular la energía cinética del satélite en la órbita, la energía total (mecánica) del satélite en la órbita, el peso del satélite en la órbita, comparar el valor con el obtenido si usamos la expresión $P = m \cdot g$ e imaginarse que el satélite pasa, sin rozamiento, a otra órbita menor a 600 km de la superficie terrestre para calcular el porcentaje de variación de la velocidad e indica si aumenta o disminuye en la nueva órbita teniendo en cuenta para todos los apartados anteriores que un satélite meteorológico de 500 kg de masa describe una trayectoria circular a una altura de 800 km sobre la superficie terrestre y a una velocidad de 7452 m/s. En esta prueba se observa que el modelo resuelve los tres primeros

apartados sin complicaciones hasta que tiene que calcular el porcentaje pedido en el último apartado. Para este apartado necesita dos *prompts* adicionales debido a que comete un uso incorrecto de los cálculos para estipular el cambio en la velocidad y no calcula bien la variación de velocidad.

En el sexto ejercicio, el modelo debe realizar para dos masas puntuales m_1 = 700 kg y m_2 = 100 kg que se encuentran fijas en los puntos A (-10,10) m y B (10,10) m, respectivamente, un esquema de la situación dibujando el vector campo gravitatorio que crea cada masa en el origen de coordenadas (0,0), calcular el campo gravitatorio (módulo, dirección y sentido) en el origen, calcular el potencial gravitatorio en el origen, calcular el trabajo necesario para llevar una masa m_3 = 200 g desde el origen hasta el punto C (10,-10) m y explicar el significado del signo del trabajo obtenido en el apartado d. En este problema se observa que tiene más complicaciones a la hora de resolver el ejercicio, por lo que necesita dos *prompts* adicionales en el que se le indica que el ejercicio está mal y en el que se le proporciona los resultados correctos para que identifique su error. En esta última interrail se observa que la IA llega al valor correcto y corrige sus errores previos (errores en la magnitud del campo gravitatorio, los componentes del campo gravitatorio y en el cálculo del potencial y el trabajo).

En el séptimo ejercicio, la IA debe calcular el módulo del momento angular de la Tierra respecto del Sol en el punto 1 si sabemos que pasa por ese punto a una velocidad aproximada de 29 740 m/s, indicar cuál es la dirección y sentido del vector momento angular de la Tierra respecto del Sol y razonar qué valor tendrá el momento angular cuando la Tierra pase por el punto 2 sabiendo que la Tierra se mueve alrededor del Sol describiendo una órbita elíptica e imaginando que no hay cuerpos celestes cercanos y que el Sol está fijo en uno de los focos de la elipse. elipse. En este problema se observa que la herramienta no tiene complicaciones a la hora de resolver los diferentes apartados y realiza un razonamiento correcto de estos.

En el octavo ejercicio, la herramienta debe realizar un dibujo de la situación indicando los vectores unitarios y las fuerzas que se ejercen sobre la masa m_3 en el punto y determina la fuerza (vector y módulo) que se ejerce sobre la masa m_3 en el punto C sabiendo que dos masas puntuales, m_1 y m_2, de 5,0 kg y 2,0 kg están situadas en los puntos fijos A (-2, 2) m y B (2, 2) m respectivamente y que una tercera masa, m_3, de 0,5 kg, se deja en libertad y en reposo en el punto C (2,-1) m. Nuevamente, se observa que el modelo no muestra complicaciones en llegar a los resultados correctos.

En el noveno ejercicio, el modelo debe calcular los siguientes apartados a partir de una ecuación de una onda armónica: el sentido de propagación de la onda en el eje x, la amplitud de la onda, la frecuencia angular o pulsación, el número de onda, la fase inicial, la fase en el instante t = 1 s y posición de la cuerda x = 0,5 m, la velocidad de propagación de la onda, la longitud de onda, el periodo y la frecuencia, la mínima distancia entre dos puntos que estén en fase, la mínima distancia entre dos puntos que estén en oposición, la máxima velocidad de cualquier partícula de la cuerda, la diferencia de fase entre dos puntos separados 0,25 m y la aceleración transversal del punto de la cuerda situado en x = 1 m en el instante t = 10 s. En este ejercicio, el modelo comete un error en el cálculo del número de onda y en un número implicado en el cálculo de la fase inicial. Estos errores desencadenan en los siguientes apartados, por lo que no llega al valor correcto de la mayoría de los apartados. Por ello, se le introduce un *prompt* adicional indicando el error. En ese instante es capaz de corregir todos los fallos de los demás apartados, llegando así a completar el ejercicio correctamente.

En el décimo ejercicio, este modelo debe hallar el periodo, la longitud de onda y la expresión matemática de la onda, la diferencia de fase entre dos puntos separados una distancia de 0,0165 m y la velocidad de vibración del origen en $t = 5 \cdot 10^{-4}$ s sabiendo que una onda sonora, que suponemos armónica, se propaga con una velocidad de 330 m/s en el sentido positivo del eje X, la frecuencia es de 10 kHz y que la amplitud de oscilación de las moléculas del medio es de $2 \cdot 10^{-4}$ m. Además, en el instante inicial, la elongación en el origen es de $1 \cdot 10^{-4}$ m, y su velocidad negativa. En este último problema se observa que la herramienta necesita dos *prompts* adicionales para llegar al resultado correcto del ejercicio debido a que en un principio no hizo bien el cambio de unidades de algunas de las variables, afectando así a la resolución de todos los apartados y posteriormente, tuvo fallos en los cálculos matemáticos, la evaluación incorrecta de las fase y cálculo de velocidad y un fallo en la expresión matemática de la onda . No obstante, logra llegar a los resultados finales.

9.2. Discusión sobre las respuestas de ChatGPT-4o a los problemas de 2º Bachiller.

En la primera prueba se le pide al modelo realizar un dibujo de la situación con todos los detalles y calcula el vector campo gravitatorio en el punto C(4,3) m, calcular el potencial gravitatorio en los puntos C(4,3) m y D(2,2) m y calcular el trabajo necesario para llevar una masa, m_3, de 5 kg desde el punto C hasta el D y razona sobre su signo respecto a dos masas iguales, m_1 y m_2, de 100 kg están situadas en los puntos A(0,0) m y B(0,3) m. Al igual que en el modelo anterior, este realiza correctamente los dos primeros

apartados sin dificultad alguna. Sin embargo, en el último apartado plantea mal la ecuación del trabajo por un error de signos. Por ello, se le introduce un *prompt* adicional para indicar el error. Con ello, llega al resultado correcto.

En el segundo ejercicio, la IA debe calcular la velocidad orbital, el periodo de su órbita, la energía mecánica en la órbita y el trabajo mínimo que deberían realizar los motores del satélite si fuese necesario pasar a otra órbita más alejada, el doble de la primera: 2h con respecto a un satélite de 1350 kg destinado a la observación climática y oceanográfica que orbita La Tierra a una altura sobre su superficie de 660 km. En esta prueba se observa que la herramienta tiene más dificultades en llegar al resultado correcto. En primer lugar, se le debe introducir un *prompt* adicional debido a que ha realizado mal un cambio de unidades en los cálculos. En ese instante, logra llegar al resultado adecuado del primer apartado. Después de haber modificado esta corrección que desencadenaba los valores correctos de los apartados siguientes, se observa que no tiene dificultades para llegar a la solución del segundo y tercer apartado. Sin embargo, en el último apartado no llega a la resolución correcta del ejercicio. Por ende, se le introduce un nuevo *prompt*, pero no identifica su error. Por ello, se le introduce otro *prompt* en el que se le proporciona la fórmula correcta que debe utilizar debido a que la ecuación que se había empleado para el cálculo del trabajo no era adecuada. Finalmente, logra llegar a la resolución correcta del ejercicio.

En el tercer ejercicio, la herramienta debe calcular el periodo de revolución del satélite expresado en horas, la energía potencial y cinética del satélite en su órbita con respecto a un satélite que tiene una masa m = 500 kg y su órbita, supuesta circular, se encuentra a una distancia de $2{,}3 \cdot 10^4$ km de la superficie terrestre. En esta prueba se observa que la IA no tiene complicaciones en calcular el periodo de revolución del satélite expresado en horas y la energía potencial. Sin embargo, a la hora de calcular la energía cinética, este realiza un buen planteamiento del problema, pero se equivoca en el cálculo matemático de este, por lo que se le introduce un nuevo *prompt* indicando el error. En este momento es capaz de llegar a la respuesta correcta.

En el cuarto ejercicio, el modelo debe hallar la velocidad de escape desde la superficie y a qué altura, desde la superficie, alcanzará un objeto si se lanza con la mitad de la velocidad de escape, sabiendo que Marte (supuesto esférico) tiene un radio de $R = 3{,}39 \cdot 10^6$ m, y el valor de la gravedad en su superficie es g = 3,71 m/s^2. En esta prueba se observa que la IA llega al resultado correcto de la velocidad sin dificultades, pero a la hora de calcular la altura comente un error en el cálculo matemático. Por ende, se le introduce un nuevo *prompt* adicional en el que se le indica el error y posteriormente llega al valor correcto del apartado.

En el quinto ejercicio, la herramienta debe calcular la energía cinética del satélite en la órbita, la energía total (mecánica) del satélite en la órbita, el peso del satélite en la órbita, comparar el valor con el obtenido si usamos la expresión $P = m \cdot g$ e imaginarse que el satélite pasa, sin rozamiento, a otra órbita menor a 600 km de la superficie terrestre para calcular el porcentaje de variación de la velocidad e indica si aumenta o disminuye en la nueva órbita teniendo en cuenta para todos los apartados anteriores que un satélite meteorológico de 500 kg de masa describe una trayectoria circular a una altura de 800 km sobre la superficie terrestre y a una velocidad de 7452 m/s. En esta prueba se observa que el modelo resuelve todos los apartados sin complicaciones.

En el sexto ejercicio, el modelo debe realizar para dos masas puntuales $m_1 = 700$ kg y $m_2 = 100$ kg que se encuentran fijas en los puntos A (-10,10) m y B (10,10) m, respectivamente, un esquema de la situación dibujando el vector campo gravitatorio que crea cada masa en el origen de coordenadas (0,0), calcular el campo gravitatorio (módulo, dirección y sentido) en el origen, calcular el potencial gravitatorio en el origen, calcular el trabajo necesario para llevar una masa $m_3 = 200$ g desde el origen hasta el punto C (10,-10) m y explicar el significado del signo del trabajo obtenido en el apartado d. En este ejercicio se observa que el modelo llega al resultado correcto de los tres primeros apartados, aunque no detalla cómo. Al llegar al penúltimo apartado (cálculo del trabajo), no muestra el valor correcto, por lo que se le añade un *prompt* adicional. En ese instante, es capaz de corregir su error (una mala interpretación del campo gravitatorio) y llegar al resultado correcto.

En el séptimo ejercicio, la IA debe calcular el módulo del momento angular de la Tierra respecto del Sol en el punto 1 si sabemos que pasa por ese punto a una velocidad aproximada de 29 740 m/s, indicar cuál es la dirección y sentido del vector momento angular de la Tierra respecto del Sol y razonar qué valor tendrá el momento angular cuando la Tierra pase por el punto 2 sabiendo que la Tierra se mueve alrededor del Sol describiendo una órbita elíptica e imaginando que no hay cuerpos celestes cercanos y que el Sol está fijo en uno de los focos de la elipse. En este ejercicio se observa que la herramienta no tiene complicaciones a la hora de resolver los diferentes apartados y realiza un razonamiento correcto de estos.

En el octavo ejercicio, la herramienta debe realizar un dibujo de la situación indicando los vectores unitarios y las fuerzas que se ejercen sobre la masa m_3 en el punto y determina la fuerza (vector y módulo) que se ejerce sobre la masa m_3 en el punto C sabiendo que dos masas puntuales, m_1 y m_2, de 5,0 kg y 2,0 kg

están situadas en los puntos fijos A (-2, 2) m y B (2, 2) m respectivamente y que una tercera masa, m_3, de 0,5 kg, se deja en libertad y en reposo en el punto C (2,-1) m. Nuevamente, se observa que el modelo no muestra complicaciones en llegar a los resultados correctos.

En el noveno ejercicio, el modelo debe calcular los siguientes apartados a partir de una ecuación de una onda armónica: el sentido de propagación de la onda en el eje x, la amplitud de la onda, la frecuencia angular o pulsación, el número de onda, la fase inicial, la fase en el instante t = 1 s y posición de la cuerda x = 0,5 m, la velocidad de propagación de la onda, la longitud de onda, el periodo y la frecuencia, la mínima distancia entre dos puntos que estén en fase, la mínima distancia entre dos puntos que estén en oposición, la máxima velocidad de cualquier partícula de la cuerda, la diferencia de fase entre dos puntos separados 0,25 m y la aceleración transversal del punto de la cuerda situado en x = 1 m en el instante t = 10 s. Aparentemente, este es un ejercicio sencillo con distintos apartados y en la mayoría de apartados la resolución es correcta. No obstante, se equivoca en el número de onda (k), y por tanto en la resolución de la longitud de onda, la velocidad de onda y otras distancias relacionadas. Por ende, se le introduce un nuevo *prompt* adicional para llegar al resultado correcto. Sin embargo, comete otros errores como el fallo en el cálculo de la fase inicial que desencadena en resultados incorrectos en el cálculo de la aceleración trascendental. En consecuencia, se le introduce otro *prompt* marcando el error y esta herramienta logra llegar al planteamiento adecuado, pero a un cálculo matemático erróneo. Entonces, se le introduce un tercer *prompt* adicional y finalmente es capaz de llegar al resultado correcto.

En el décimo ejercicio, este modelo debe hallar el periodo, la longitud de onda y la expresión matemática de la onda, la diferencia de fase entre dos puntos separados una distancia de 0,0165 m y la velocidad de vibración del origen en $t = 5 \cdot 10^{-4}$ s sabiendo que una onda sonora, que suponemos armónica, se propaga con una velocidad de 330 m/s en el sentido positivo del eje X, la frecuencia es de 10 kHz y que la amplitud de oscilación de las moléculas del medio es de $2 \cdot 10^{-4}$ m. Además, en el instante inicial, la elongación en el origen es de $1 \cdot 10^{-4}$ m, y su velocidad negativa. En este último problema se observa que la herramienta necesita dos *prompts* adicionales para llegar al resultado correcto del ejercicio debido a que en un principio no hizo bien el cambio de unidades de algunas de las variables, afectando así a la resolución de todos los apartados y posteriormente, tuvo fallos en los cálculos numéricos. No obstante, logra llegar a los resultados finales.

9.3. Discusión de la comparación de la performance de todos los modelos.

Tabla 8

Nota media de cada modelo en cada una de las pruebas.

Desempeño Global de cada modelo						
		ChatGPT-3,5	**ChatGPT-4**	**ChatGPT-4o**	**ChatGPT-4o en You.com**	**ChatGPT-o1 preview**
Filosofía	**1º Bachiller 1**	5,50	9,40	9,20		
	Disertación	4,00	5,50	5,75		
	EVAU		6,14	7,56		
Física	**Problema 1 (3º ESO)**	3,33	10,00	10,00	9,17	10,00
	Problemas 2, 3, 4, 5, 6, 7 , 8, 9 (3º ESO)		8,15	8,15	7,50	7,50
	Problemas 2º Bachiller (10 problemas)		8,25	7,75		

esto le preguntamos a iris luego que he visto una cosa

- **Exámenes de Filosofía**

En la tabla se puede ver como el desempeño de esta IA ha mejorado con el tiempo. El modelo más reciente, ChatGPT-4o, es el que obtiene mejores resultados en casi todas las pruebas. Su desempeño global es el mejor exceptuando el primer examen de primero de Bachiller.

Los tres modelos muestran una alta capacidad a la hora de extraer, sintetizar o explicar un texto. Es decir, poseen una gran comprensión del lenguaje.

En la comparación de los tres primeros modelos en el primer examen se pudo concluir que ChatGPT-3.5 presenta dificultades en el razonamiento abstracto y contextual que requiere pensar más allá del texto y realizar conexiones más complejas entre conceptos, a diferencia de sus dos modelos sucesores.

En la disertación se observa cómo los tres modelos tienen un rendimiento negativo. ChatGPT-3.5 obtiene un 4, ChatGPT-4 un 5,5 y ChatGPT-4o un 5,75. Esto es calificándolos con la rúbrica propuesta para los alumnos de 2º de Bachiller. Se puede ver que el desempeño es malo en los tres modelos porque tienden a limitarse a copiar o parafrasear ideas del texto base proporcionado como apoyo en lugar de desarrollar argumentos originales. Esto muestra una falta de argumentación propia y reflexión crítica. Además, a diferencia de ChatGPT-4, los dos primeros modelos analizados no dan un contraargumento. En síntesis, el tratamiento que los tres modelos dan a la disertación se asemeja más al de un comentario de texto, pues se centran en analizar el contenido del texto base en lugar de generar una argumentación estructurada y fundamentada. Dicho esto, la estructura también es confusa en el primer modelo y mejora en los dos segundos.

En los exámenes de EVAU solo se analiza a ChatGPT-4 y ChatGPT-4o. Sus medias son de 6,14 y 7,56 respectivamente. La diferencia de un punto se debe principalmente a que el segundo modelo explica con mayor claridad los conceptos y, en ocasiones, también profundiza mejor. Es destacable que, a veces, el modelo obtiene mejor calificación que el primero al captar más ideas fundamentales sin utilizar necesariamente una extensión más larga, lo cual demuestra que tiene una capacidad para explicar de forma precisa y concisa los conceptos.

- **Exámenes de física:**

Las conclusiones extraídas de los resultados de física utilizando ChatGPT-3.5 no han sido muy satisfactorias. Esta herramienta no muestra mucha capacidad en la comprensión a la hora de analizar los enunciados de los problemas y comete muchos errores que tienen que ser corregidos por el propio usuario. Carece de habilidad y precisión en las respuestas.

A diferencia del modelo anterior, las respuestas obtenidas por las distintas versiones de ChatGPT-4 (ChatGPT-4.o, ChatGPT o1-preview) en cuanto a los exámenes de cinemática de 3ºESO son mucho más satisfactorias. En la mayoría de los ejercicios llega a las respuestas correctas sin necesidad de introducir más *prompts* para aclarar confusiones y sin fallos. Si bien es cierto que llama la atención que absolutamente todos los modelos tanto en OpenAI como en you.com, en el caso de ChatGPT-4.o, no consiguen llegar a las respuestas correctas del sexto ejercicio al mostrar cierta confusión con la manera de interpretar el problema. En concreto, el apartado que llama más la atención es el del cálculo de la aceleración final del segundo intervalo, ya que la mayoría de estos modelos no consideran que la aceleración final del segundo intervalo es la aceleración inicial del tercer intervalo y por consiguiente todos afirman que es 0 m/s^2 hasta que se les introduce nuevos *prompts* para clarificar la situación física de la partícula. Sin embargo, los resultados en general han sido exactamente como los esperábamos al inicio del proyecto. El modelo de pago da mejores resultados que la versión anterior en cuanto a la resolución de *inverse problems*.

En cuanto a las respuestas de ChatGPT-4o ante los enunciados según el idioma en el que se introducen los *prompts* (inglés y español), destaca una discrepancia de respuestas en el quinto ejercicio realizado por you.com, sin embargo, el resto de ejercicios proporcionan los mismos resultados tanto en inglés como español, por lo que esta alternancia de resultados podría deberse a un saturación de la herramienta o simplemente a un fallo puntual por la IA, ya que durante el proceso de examinación de los ejercicios de física, la herramienta accesible en you.com presentaba una notable lentitud de respuestas y en numerosas ocasiones el programa quedaba congelado. Ante este problema, se puede concluir que a pesar de que las respuestas proporcionadas en you.com y OpenAI frecuentemente son las mismas, la calidad de este primer programa al realizar gráficas es pobre y su lentitud es incómoda a la hora de utilizarlo. Por lo que OpenAI es más manejable.

Con respecto a los problemas extraídos de la unidad de interacción gravitatoria incluida en el temario de física en la selectividad, se observa que tanto ChatGPT-4 como ChatGPT-4o, solucionan los ejercicios de manera parecida, siendo el ChatGPT-4 un poco más eficaz. No obstante, al aumentar el nivel de dificultad de

los exámenes, ambos modelos presentan con frecuencia más errores en los cálculos matemáticos y en la interpretación del contexto que en el planteamiento.

10. Limitaciones y perspectivas futuras

A lo largo de la realización del proyecto se encuentran varias limitaciones como el tiempo, el avance rápido en la materia o la propia muestra en los formularios. A continuación, se detalla algunas de estas limitaciones y sugerencias a futuros trabajos que pueden llevarse a cabo para afrontarlas.

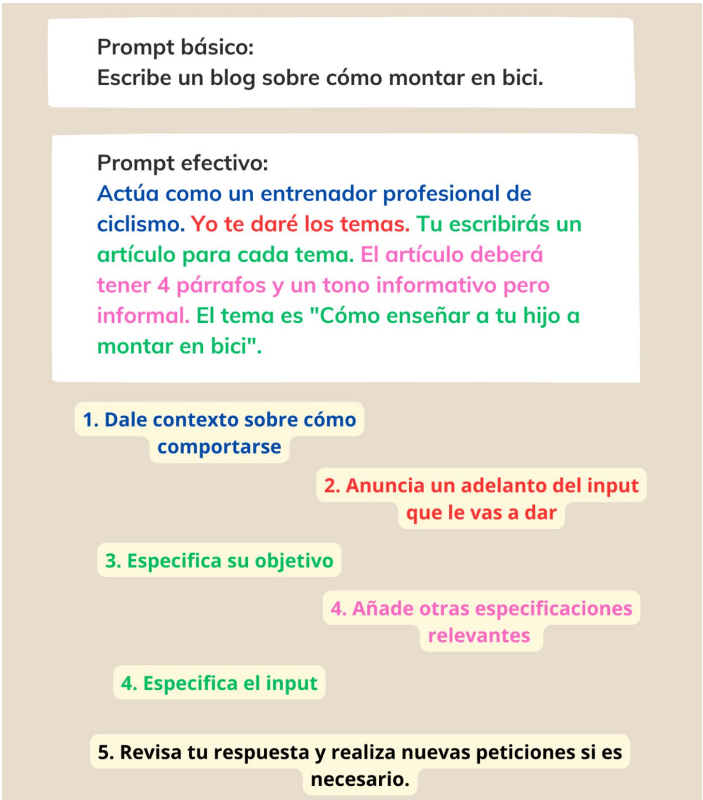
Debido al rápido avance del sector tecnológico durante los últimos dos años, se han observado limitaciones en la evaluación de los modelos. En este estudio, se planteaba en un inicio comparar solamente al modelo ChatGPT-3.5 junto a su versión de pago, ChatGPT-4. Sin embargo, conforme avanzaba el tiempo nuevos modelos surgían y se decidió incluirlos también. Asimismo, la versión de ChatGPT-3.5 fue eliminada del alcance público a partir del 19 de julio de 2024 (AndroidSage, 2024). Este hecho imposibilitó que se pudieran realizar más pruebas a este modelo, lo que limita la muestra que se tiene frente al resto. Asimismo, debido a estos factores el tiempo jugó un papel limitante, puesto que a medida que estos modelos surgían se requería tiempo de estudio para conocer las diferencias frente a los modelos previos y para realizarles a ellos también la evaluación. Por ello, sería interesante plantear trabajos que evalúen a los diferentes modelos con una cantidad mayor de exámenes.

Además, en este estudio solo se le evalúan las asignaturas de física y de filosofía en determinados niveles. Sería interesante replicar este estudio en niveles superiores universitarios o en áreas de conocimiento diferente. Otro estudio interesante podría emplear a profesores expertos en las materias para evaluar a los modelos y así comparar y usar las calificaciones que estos asignasen.

Por otro lado, sería interesante plantear futuros estudios que analicen la efectividad, en mayor profundidad, de estas IA en relación al uso de *prompts*. Esta es una variable que consiste en la combinación de palabras usada para cuestionar al modelo. Al emplear técnicas de ingeniería de *prompts*, los usuarios pueden orientar a ChatGPT para que genere respuestas más precisas, relevantes y útiles. Es decir, cuanto más clara, con más contexto o más específica sea la pregunta se pueden asegurar mejores resultados tal y como se observa en la figura 16.

Figura 16

Instrucciones para realizar un prompt efectivo



Fuente: elaboración propia.

Respecto a los formularios, estos ofrecen una reflexión sobre los hábitos de uso y la percepción de estudiantes y docentes, pero surge la pregunta de si los resultados pueden ser transferidos a una muestra más

amplia. Actualmente, existen muy pocos estudios que analicen la percepción de los docentes (Lindner et al., 2019), y aún menos la de los estudiantes, a pesar de que ellos constituyen el foco central del ámbito educativo.

Según Lindner et al. (2019), el conocimiento de los docentes sobre inteligencia artificial está influenciado por temas mediáticos actuales y perciben una falta de materiales didácticos adecuados y ejemplos prácticos en este campo. Por ello, sería útil realizar la guía con una mayor profundidad para poder brindar una visión más holística a los profesores en la implementación de IA educativa.

11. Conclusiones

Tras llevar a cabo los objetivos que se establecieron al inicio del proyecto y plantear diversas hipótesis sobre las respuestas de las encuestas proporcionadas al profesorado y al alumnado se concluyen varias cosas.

En la primera hipótesis formulada, se planteó que los docentes harían más uso de la IA que el alumnado. Al realizar el tratamiento estadístico de la muestra se observa que la hipótesis resulta incorrecta, no existen diferencias significativas entre el uso de la IA por parte de alumnado y profesorado, existiendo outliers de alumnado que la usa a diario, este hallazgo refuta también la hipótesis 4, ya que la edad no supone ninguna diferencia en el uso de la IA. Asimismo se puede constata que la hipótesis dos en la que se enuncia que en el cuerpo docente se usan más las herramientas generativas de lenguaje entre aquellos con menos años de experiencia, queda refutada.

Con respecto a la suposición de una distribución uniforme del uso de la IA según la variable género si bien es cierto que no existen diferencias significativas entre el uso de la IA por parte de aquellas personas que se identifican con un género binario (masculino o femenino) las que se identifican con otro género o prefieren no dar información sobre el mismo reportan significativamente un mayor uso de ChatGPT.

Finalmente, la quinta hipótesis («los alumnos preferirán el apoyo de un profesor que el de la IA.»), también queda verificada debido a que, la gran mayoría prefiere el trato con un profesor.

En cuanto a la evaluación de los diferentes modelos, en las pruebas de filosofía se puede observar como la hipótesis sobre una mejora significativa en las respuestas de la herramienta en las versiones más recientes se cumple. Corroboramos esta hipótesis puesto que ChatGPT 4o es el que obtiene la nota más alta a lo largo de todas las pruebas.

También se planteó como hipótesis que la eficacia del modelo Chat GPT-3.5 en la generación de respuestas atendiendo al número de revisiones del prompt sería significativamente menor en preguntas de filosofía que en los problemas de física. Esto se cumple, puesto que no logra resolver el primer problema de física a pesar de introducir numerosas instrucciones. En filosofía, siempre construye algún texto.

.

Bibliografía y webgrafía

ACUÑA, C. C. (2024). Cómo usar la Inteligencia Artificial para la consecución de los ODS de la Agenda 2030: una visión desde la ética, la innovación y los algoritmos verdes. Consultor de los ayuntamientos y de los juzgados: Revista técnica especializada en administración local y justicia municipal, (9), 21. https://prodigital.web.uclm.es/wp-content/uploads/2024/09/Como_usar_la_Inteligencia_Artificial_para_la_cons.pdf

Aguilar, S. J., Swartout, W., Nye, B., Sinatra, G. M., Wang, C., & Bui, E. (2024). Critical Thinking and Ethics in the Age of Generative AI in Education. USC Center for Generative AI and Society. https://doi.org/10.35542/osf.io/7dr9j

AndroidSage. (2024b, julio 19). ChatGPT 3.5 is Discontinued in Support of GPT-4o mini - Android Sage. Android Sage. https://www.androidsage.com/2024/07/19/chatgpt-3-5-is-discontinued-in-support-of-gpt-4o-mini/?utm_source=chatgpt.com

Ansede, M., Ansede, M., & Ansede, M. (2024, 25 abril). El exceso de palabras como “encomiable” y “meticuloso” sugiere el uso de ChatGPT en miles de estudios científicos. El País. https://elpais.com/tecnologia/2024-04-25/el-exceso-de-palabras-como-encomiable-y-meticuloso-sugiere-el-uso-de-chatgpt-en-miles-de-estudios-cientificos.html?utm_source=chatgpt.com

Ayala Ayala, L. K. (2021). Análisis de las estrategias de adaptación y mitigación relacionadas al cambio climático frente al ODS N° 13" acción por el clima" presentes en el PDM de Bucaramanga. http://repositorio.uts.edu.co:8080/xmlui/handle/123456789/7508

Ayuso, A. (2024, 9 septiembre). El ChatGPT educativo que ayuda a los profesores a preparar sus clases y aligerar las «cargas administrativas» para evitar que se quemen. El Periódico de España. https://www.epe.es/es/reportajes/20240908/chatgpt-educativo-profesores-inteligencia-artificial-107745578?utm_source=chatgpt.com

Carrillo M-Feduchi, G. (2022). Métricas de eficiencia en los data centers (PUE, CUE, WUE). DataCenterDynamics. https://www.datacenterdynamics.com/es/features/m%C3%A9tricas-de-eficiencia-en-los-data-centers-pue-cue-wue/

Cazzaniga, M., Jaumotte, M. F., Li, L., Melina, M. G., Panton, A. J., Pizzinelli, C., Rockall, E., & Tavares, M. M. M. (2024). Gen-AI: Artificial intelligence and the future of work. International Monetary Fund.

Codina, L. (2023). Buscadores alternativos a Google con IA generativa: análisis de You.com, Perplexity AI y Bing Chat [Alternative search engines to Google with generative AI: analysis of You.com, Perplexity AI and Bing Chat]. INFONOMY, 1. https://doi.org/10.3145/infonomy.23.002

Crawford, K. (2024). Generative AI’s environmental costs are soaring — and mostly secret. Nature, 626(8000), 693. https://doi.org/10.1038/d41586-024-00478-x

Diego Olite, F., Morales Suárez, I., & Vidal Ledo, M. (2023). Chat GPT: origen, evolución, retos e impactos en la educación. Educación Médica Superior, 37(2).

Elliott, D., & Soifer, E. (2022b). AI Technologies, Privacy, and Security. Frontiers In Artificial Intelligence, 5. https://doi.org/10.3389/frai.2022.826737

Estrategia Andaluza de Inteligencia Artificial 2030. (s. f.). Junta de Andalucía. https://www.juntadeandalucia.es/organismos/ada/estructura/transparencia/planificacion-evaluacion-estadistica/planes/detalle/427612.html

Europa Press. (2023). Una consulta en ChatGPT consume tres veces más energía que en el buscador de Google. https://www.europapress.es/portaltic/sector/noticia-consulta-chatgpt-consume-tres-veces-mas-energia-buscador-google-20230728164651.html

Habib, S., Vogel, T., Anli, X., & Thorne, E. (2023). How does generative artificial intelligence impact student creativity? Journal Of Creativity, 34(1), 100072. https://doi.org/10.1016/j.yjoc.2023.100072

Haraway, D. J. (1995). Ciencia, cyborgs y mujeres: la reinvención de la naturaleza (Vol. 28). Universitat de València.

Holmes, W., Hui, Z., Miao, F., & Ronghuai, H. (2021). Inteligencia artificial y educación: Guía para las personas a cargo de formular políticas. Publicaciones UNESCO.

Horne, B. D., Nevo, D., O’Donovan, J., Cho, J., & Adalı, S. (2019). Rating Reliability and Bias in News Articles: Does AI Assistance Help Everyone? Proceedings Of The International AAAI Conference On Web And Social Media, 13, 247-256. https://doi.org/10.1609/icwsm.v13i01.3226

Li, P., Yang, J., Islam, M. A., & Ren, S. (2023). Making AI Less «Thirsty»: Uncovering and Addressing the Secret Water Footprint of AI Models. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2304.03271

Lindner, A., & Romeike, R. (2019). Teachers' perspectives on artificial intelligence. En E. Jasutė & S. Pozdniakov (Eds.), ISSEP 2019 - 12th International Conference on Informatics in Schools: Situation, Evaluation and Perspectives, Local Proceedings (pp. 22-29). Larnaca, Chipre.

Lo, C. K. (2023). What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. Education Sciences, 13(4), 410. https://doi.org/10.3390/educsci13040410

Luckin, R., & Holmes, W. (2016). Intelligence unleashed: An argument for AI in education.

Machine Learning in the Real World | Blog. (s. f.). https://www.insiris.com/blog/machine-learning-in-the-real-world

Markey, E. J. (2024, 1 de febrero). Markey, Heinrich, Eshoo, Beyer introduce legislation to investigate, measure environmental impacts of artificial intelligence. Edward Markey. https://www.markey.senate.gov/news/press-releases/markey-heinrich-eshoo-beyer-introduce-legislation-to-investigate-measure-environmental-impacts-of-artificial-intelligence

mundoestudiante. (2024, 4 enero). 【Exámenes Selectividad Resueltos】 EVAU Corregidos últimos años. Mundoestudiante. https://www.mundoestudiante.com/examenes-resueltos/selectividad/

OpenAI. (2024). ChatGPT. OpenAI. https://www.openai.com

Orrù, G., Piarulli, A., Conversano, C., & Gemignani, A. (2023). Human-like problem-solving abilities in large language models using ChatGPT. Frontiers In Artificial Intelligence, 6. https://doi.org/10.3389/frai.2023.1199350

Palumbo, D. (2024, 23 abril). The A.I. lie. Muddy Colors. https://www.muddycolors.com/2024/04/the-a-i-lie/?utm_source=chatgpt.com

Parra, S. (2024, 10 octubre). La revolución de la IA en la medicina: AlphaFold y el Nobel de Química 2024. National Geographic España. https://www.nationalgeographic.com.es/ciencia/revolucion-ia-medicina-alphafold-y-nobel-quimica-2024_23415

Patel, R. (2024, 12 noviembre). LLM Benchmarks: A Comprehensive Guide to AI Model Evaluation | PromptLayer. PromptLayer. https://blog.promptlayer.com/llm-benchmarks/?utm_source=chatgpt.com

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). Carbon Emissions and Large Neural Network Training. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2104.10350

Pistilli, G. (2022). What lies behind AGI: Ethical concerns related to LLMs. Revue Éthique et Numérique. https://hal.science/hal-03607808

Prasad, A., Segarra, P., & Villanueva, C. E. (2020). Situating knowledges through feminist objectivity in organization studies. En Routledge eBooks (pp. 73-88). https://doi.org/10.4324/9780429279720-6

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Of Things And Cyber-Physical Systems, 3, 121-154. https://doi.org/10.1016/j.iotcps.2023.04.003

Rivera Parra, Á. A. (2023, 28 de agosto). Automatización de marcado de texto en TEI. Repositorio Institucional Séneca. http://hdl.handle.net/1992/70089

Rouhiainen, L. (2018). Inteligencia artificial: 101 cosas que debes saber hoy sobre nuestro futuro. Madrid: Alienta Editorial.

Sistema de Información Energética. (2024, octubre). Consumo energético del Gobierno de Navarra. Gobierno de Navarra.

Temsah, M., Jamal, A., Alhasan, K., Temsah, A. A., & Malki, K. H. (2024). OpenAI o1-Preview vs. ChatGPT in Healthcare: A New Frontier in Medical AI Reasoning. Cureus. https://doi.org/10.7759/cureus.70640

UNESCO IESALC. (2023). ChatGPT e Inteligencia Artificial en la Educación Superior: Guía de Inicio Rápido. UNESCO IESALC. https://www.unesco.org/open-access/terms-use-ccbysa-sp.

UNESCO. (2019). Consenso de Beijing sobre la inteligencia artificial y la educación.

ANEXOS

Anexo A: Datos de los formularios.

Anexo B: Respuestas a todos los exámenes.

Anexo C: Rúbricas y evaluación de los exámenes.

Anexo D: Guía de uso.