Como citar este artículo: R. Rodriguez-Echeverría, J. D. Gutiérrez, J. M. Conejero and Á. E. Prieto, "Analysis of ChatGPT Performance in Computer Engineering Exams," in IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, doi: 10.1109/RITA.2024.3381842.

Análisis del desempeño de ChatGPT en exámenes de Ingeniería Informática

Roberto Rodríguez-Echeverría, Juan D. Gutiérrez, José M. Conejero y Álvaro E. Prieto

***Abstract*— La aparición de ChatGPT a finales del 2022 supuso un hito en el campo de las Inteligencias Artificiales Generativas, pero también causó un gran revuelo en el mundo académico. Por primera vez, una interfaz sencilla permitía a cualquier persona acceder a un modelo de lenguaje de gran tamaño y utilizarlo para generar texto. Estas capacidades pueden tener, sin duda, un impacto relevante en las metodologías de enseñanza-aprendizaje y también en los métodos de evaluación. Con el objetivo de obtener una medida real del posible desempeño de ChatGPT en la resolución de exámenes, se le ha puesto a prueba con los exámenes de 15 asignaturas de Ingeniería del Software de un grado de Ingeniería Informática. A la luz de los resultados, ChatGPT es capaz de lograr un desempeño relevante en estos exámenes; ya que, es capaz de superar una cantidad significativa de preguntas y problemas de diferente naturaleza en múltiples asignaturas. Como contribución fundamental, se proporciona un estudio detallado de los resultados por tipología de preguntas y problemas, que permite establecer unas recomendaciones a tener en cuenta en el diseño de los métodos de evaluación. Además, se presenta un análisis del impacto del aspecto no determinista de ChatGPT en las respuestas de las preguntas de test.**

***Index Terms*— Inteligencia artificial, ChatGPT, experimento, examen, evaluación.**

I. Introducción

La influencia de la Inteligencia Artificial (IA) en la docencia de Ingeniería Informática estaba clara ya desde finales del siglo XX. En la edición de 1997 de las Jornadas sobre la Enseñanza Universitaria de la Informática (JENUI) [1], el 25 % de las ponencias la incluían directamente en su título. En ellas se compartían con la comunidad educativa las diferentes formas en las que esta disciplina estaba entrando en los programas educativos. Tan solo un cuarto de siglo después, la situación ha cambiado tanto que el tema a tratar es qué consecuencias tendrá el uso de la IA en todos los ámbitos de la educación superior. Por ejemplo, en [2] se propone la utilización de modelos de IA para asistir en la evaluación de trabajos informáticos complejos.

En 1943, McCulloch y Pitts presentaron el perceptrón [3], poniendo en marcha un campo del conocimiento con un potencial inmenso. En una elipsis narrativa digna de Kubrick, este trabajo permitió que la empresa OpenAI presentara a finales del 2022 ChatGPT [4], una interfaz para el acceso a su modelo del lenguaje de gran tamaño (*Large Language Model*, LLM) GPT 3.5. Dicho logro no hubiese sido posible sin la presentación del *transformer* [5], un modelo de aprendizaje profundo presentado por Google en 2017, basado en el concepto de atención, que ha demostrado ser fundamental en el campo de los LLM.

Un LLM es un modelo de IA que ha sido entrenado utilizando grandes corpus de texto. Estos modelos utilizan técnicas de aprendizaje profundo y son capaces de generar texto que se asemeje al humano. Algunos ejemplos de LLM son GPT-3 de OpenAI [6], OPT de Meta [7] o BLOOM [8]. Este último tiene la peculiaridad de tratarse de una alternativa de acceso abierto (*open-access*) en todos sus aspectos, mientras que los demás son desarrollos privativos. Además, estos modelos son capaces de realizar satisfactoriamente tareas como la traducción automática, la generación de texto, la clasificación de texto y la respuesta a preguntas.

La evaluación de las capacidades de ChatGPT ha estado en el punto de mira desde el principio. A los dos meses de su publicación, ChatGPT ya tuvo tiempo de enfrentarse a pruebas de acceso para medicina [9] y abogacía [10]. En el primer caso, el rendimiento de ChatGPT fue comparable al de un estudiante de tercero de medicina. En el segundo caso superó el 50 % de las preguntas.

En [11] se realiza una revisión sistemática de la utilización de *chatbots* en el ámbito de la educación, analizando las áreas en las que se han utilizado, su papel pedagógico, su utilización en tareas de tutorización y su potencial en una educación personalizada. La combinación de un *chatbot* con un LLM de gran fiabilidad parece prometedora para su uso en educación.

La capacidad de ChatGPT como herramienta de ayuda en la escritura se pone a prueba en [12]. El autor se plantea como reto generar, con su ayuda, un trabajo académico. El resultado obtenido permite concluir que se trata de una herramienta útil y que aumenta la eficiencia de quienes la usan, y que será necesario encontrar nuevas formas de evaluación que se centren en los aspectos que las IA no puedan sustituir, como la creatividad y el pensamiento crítico.

Sin embargo, la mejora de las capacidades humanas utilizando este tipo de tecnologías tiene las mismas implicaciones que el uso de fármacos o sustancias estimulantes para potenciar artificialmente el rendimiento de los deportistas, algo en lo que se centra [13]. Tras analizar las ventajas e implicaciones del uso de *chatbots* en la investigación, así como sus limitaciones, el autor muestra las consideraciones éticas y los po-

Todos los autores pertenecen al Departamento de Ingeniería de Sistemas Informáticos y Telemáticos, Universidad de Extremadura, Cáceres, España, salvo Juan D. Gutiérrez, que pertenece al Departamento de Electrónica y Computación, Universidad de Santiago de Compostela, Lugo, España. Autor de correspondencia: Roberto Rodríguez-Echeverría (rre@unex.es)

sibles sesgos que tiene la utilización de tecnologías como ChatGPT en trabajos de investigación, concluyendo que este tipo de tecnologías tiene el potencial de revolucionar la investigación académica.

Por su parte, en [14] el autor determina que ChatGPT sí es capaz de exhibir rasgos creativos en sus resultados, poniendo en peligro la integridad de los exámenes online y, por tanto, su evaluación. La posibilidad de usar ChatGPT ilícitamente en los exámenes obliga a replantearse los métodos de evaluación para que sigan siendo justos para todo el alumnado.

El impacto de este tipo de modelos en la realidad de la educación en general es motivo de estudio en todo el planeta. Las dificultades de detectar y prevenir la deshonestidad académica se estudian en [15]. En este trabajo se sugieren estrategias que las universidades pueden adoptar para garantizar un uso ético y responsable de estas herramientas. Entre ellas, el desarrollo de políticas y procedimientos, la formación y el apoyo, y el uso de diversos métodos para detectar y prevenir las trampas. Con otro enfoque, la enorme aplicabilidad de herramientas de IA transformadoras como ChatGPT, haciendo hincapié en sus posibles repercusiones positivas y negativas en diversos sectores se estudia en [16]. A pesar de reconocer sus limitaciones y posibles problemas éticos, en este trabajo se tienen en cuenta las mejoras de productividad obtenidas utilizando estas tecnologías en diferentes ámbitos. Por último, quienes tengan interés en obtener una perspectiva pragmática, lejos de los sesgos inherentes a posturas extremas, sobre el desafío que enfrenta la educación encontrará en [17] una reflexión que aborda las ventajas, inconvenientes, potencialidades, límites y retos de las tecnologías generativas de inteligencia artificial en la educación.

Este trabajo es una extensión del presentado en [18], en el que se analiza el impacto de ChatGPT en los métodos de evaluación de un grado de Ingeniería Informática. Con este objetivo, se diseñó un experimento consistente en intentar superar los exámenes de 15 asignaturas dentro de la rama de Ingeniería del Software. Aunque gran parte de las asignaturas incluyen métodos de evaluación más allá de los exámenes, como por ejemplo el desarrollo de proyectos, en este primer trabajo se analizó solo su impacto en los exámenes. Como extensión, se presenta un estudio de la frecuencia de las respuestas proporcionadas por ChatGPT para las preguntas de test, con el objetivo de valorar el impacto del aspecto no determinista de esta herramienta en este tipo de estudios.

El resto del trabajo está organizado de la siguiente forma. En la Sección II se describe el proceso y las pautas seguidas en el desarrollo del experimento. Posteriormente, se analizan detalladamente los resultados en la Sección III y se presentan una serie de recomendaciones en la Sección IV. Y finalmente se presentan las conclusiones principales obtenidas y las líneas de trabajo futuro en la Sección V.

II. Metodología

En este apartado se presentan las preguntas de investigación, los principales pasos de desarrollo del experimento y las pautas seguidas para proporcionar las preguntas de examen a ChatGPT, dadas sus características. Además, se definen varias dimensiones para la categorización de las preguntas de examen para poder realizar un análisis más detallado. Finalmente, se especifica el método seguido para reducir el impacto del aspecto no determinista de ChatGPT en las respuestas de las preguntas de test.

A. Preguntas de investigación

En este trabajo se pretende dar respuesta a las siguientes preguntas:

1) ¿Es capaz ChatGPT de superar los exámenes?
2) ¿Acierta más preguntas de las que falla?
3) ¿Influye el tipo de pregunta de examen?
4) ¿Influye el tipo de aplicación del conocimiento?
5) ¿Cómo ha sido su desempeño por asignatura?

B. Pasos principales

El proceso de desarrollo del experimento presentado en este trabajo ha seguido los siguientes seis pasos, como muestra la Fig. 1.

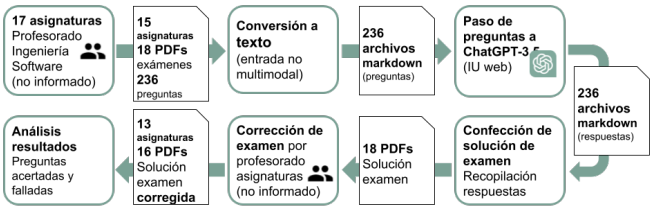

Fig. 1. Vista general del proceso seguido

1) Se ha contactado con el profesorado de todas las asignaturas de la materia de Ingeniería Software (17) para solicitarles exámenes reales y otros métodos de evaluación del curso 2021-2022. No se les ha informado del uso de ChatGPT. Solo se les informó de la intención de realizar una evaluación conjunta de los métodos de evaluación de las asignaturas de la materia. A esta solicitud respondieron 15 de las 17 asignaturas consideradas, pero solo 13 proporcionaron sus correcciones.
2) Se han organizado por asignaturas todos los exámenes proporcionados y se han descrito brevemente para indicar qué tipo de preguntas contienen (test, preguntas cortas, problemas), así como si contienen figuras o requieren alguna información de contexto.
3) Se ha generado una versión completamente textual de cada examen para que pueda ser procesado por ChatGPT. Se ha procurado que las modificaciones realizadas en cada caso sean mínimas, para que el resultado final sea lo más parecido posible al que obtendría un alumno. Estas modificaciones son de diferente índole. En algunos casos ha bastado con dividir una pregunta en diferentes partes, para que ChatGPT responda a todas ellas. En otros casos, en los que la pregunta está acompañada por una figura, ésta se ha descrito de forma textual.
4) Se ha proporcionado a ChatGPT la nueva versión del examen, conformada como una conversación siguiendo las pautas descritas más adelante. La versión de ChatGPT utilizada ha sido la publicada el 15 de diciembre del

2022[1].

5) A partir de las respuestas obtenidas de ChatGPT se ha generado un examen resuelto, que se ha enviado al profesorado de cada asignatura para su corrección.
6) Se ha realizado un análisis de los resultados obtenidos por ChatGPT en cada examen. Teniendo en cuenta la calificación obtenida, se ha analizado pregunta por pregunta el desempeño logrado y los comentarios de corrección realizados por el profesorado.

Para una información más detallada de los aspectos técnicos y organizativos de la metodología seguida, se sugiere revisar los materiales disponibles en nuestro repositorio[2].

C. Pautas de adaptación

Las pautas seguidas para proporcionar las preguntas de los exámenes a ChatGPT han sido las siguientes:

1) Las preguntas se proporcionan en español para mantener la máxima consistencia con el examen real y también para obtener las respuestas en el mismo idioma.
2) Si el examen es de tipo test y las preguntas no tienen relación entre sí, cada pregunta se realiza en una conversación aparte. De esta manera, no se genera un contexto artificial relacionado con el orden de realización de las preguntas.
3) Si el examen es de preguntas a desarrollar y presentan múltiples apartados o subpreguntas, se proporciona cada apartado por separado manteniendo la misma conversación. De esta manera se evita que la respuesta de ChatGPT pueda cortarse por ser de longitud excesiva y superar algún límite preestablecido.
4) En preguntas que incluyen código y hacen referencia a una línea concreta, se numeran las líneas de código para referirse a la línea por número.
5) En preguntas que incluyen una figura representando una estructura de datos, se proporciona una descripción textual de sus elementos principales y relaciones. Por ejemplo, en el caso de un grafo, se pueden proporcionar sus conjuntos de nodos y aristas. En algunos casos, cabe la posibilidad también de usar sintaxis textuales para definición de diagramas, como por ejemplo la sintaxis de Mermaid[3].
6) Las preguntas que contienen un contexto no explícito en el examen y desconocido para nosotros, por ejemplo las referidas a algún problema o proyecto realizado en la asignatura, se realizan sin proporcionar información adicional a ChatGPT.
7) Las tablas de datos se proporcionan según el formato CSV[4], aunque otros formatos serían posibles, por ejemplo Markdown[5].
8) En las preguntas de completar huecos, se utiliza la virgulilla (~) para indicar dónde debe ir la respuesta.

D. Categorías de preguntas

Con el fin de poder realizar un análisis detallado de los resultados obtenidos, hemos categorizado las preguntas en base a tres dimensiones: tipo de pregunta, tipo de conocimiento y tipo de aplicación.

Dentro de la primera dimensión, los tipos posibles de pregunta de examen son: pregunta de test, pregunta de teoría o problema. La pregunta de test es de respuesta única y contiene cuatro posibles respuestas. La pregunta de teoría consiste en un enunciado en el que se pide desarrollar algún contenido teórico de la asignatura. Por último, los problemas consisten en enunciados que plantean ejercicios o aplicación de contenidos de teoría a ejemplos prácticos.

En cuanto a los tipos de conocimiento, solo hemos considerado dos categorías: definición literal y definición aplicada. Estas categorías se corresponden, respectivamente, con el primer y último nivel cognitivo base de Bloom: conocimiento y aplicación. Algunos ejemplos del primer tipo son: definir un concepto, indicar un término específico o listar unas propiedades. Por su parte, el segundo tipo se refiere a la aplicación (uso) del conocimiento a un caso o ejemplo concreto especificado en el examen.

Es importante señalar que la primera y segunda dimensión son independientes entre sí, mientras que la tercera dimensión se usa para obtener una categorización de grano más fino para las preguntas clasificadas como definición aplicada.

Finalmente, en cuanto al tipo de aplicación, hemos considerado los 9 tipos siguientes:

1) *Análisis de lenguajes de programación* (LPA). La pregunta contiene un fragmento de código que se tiene que leer y entender para poder contestar la pregunta.
2) *Generación de lenguajes de programación* (LPG). La pregunta solicita una respuesta en código fuente. Por ejemplo, escribe una consulta SQL determinada.
3) *Semántica operacional* (SO). La pregunta contiene un fragmento de código cuya ejecución debe entenderse para poder contestar la pregunta.
4) *Cálculo matemático* (CA). La pregunta requiere realizar algún tipo de cálculo matemático explícito o implícito.
5) *Algoritmo o método* (AM). La pregunta requiere seguir un algoritmo o un método con múltiples pasos.
6) *Análisis de expresiones algebraicas* (AEA). La pregunta contiene expresiones algebraicas que se tienen que leer y entender para poder contestar la pregunta.
7) *Generación de expresiones algebraicas* (GEA). La pregunta solicita una respuesta en forma de expresión algebraica.
8) *Análisis de diagrama* (AD). La pregunta contiene un diagrama que representa, por ejemplo, una estructura de datos o un modelo y que se tiene que leer y entender para poder contestar la pregunta.
9) *Generación de diagrama* (GD). La pregunta solicita una respuesta en forma de diagrama.

Las preguntas de los exámenes pueden pertenecer a varios tipos de aplicación a la vez. Por ejemplo, una pregunta que presenta un fragmento de código incompleto puede solicitar

[1] https://help.openai.com/en/articles/6825453-chatgpt-release-notes
[2] https://github.com/i3uex/jenui23_chatgpt
[3] https://github.com/mermaid-js/mermaid
[4] https://datatracker.ietf.org/doc/html/rfc4180
[5] https://en.wikipedia.org/wiki/Markdown

generar el código que falta (análisis y generación de lenguajes de programación).

E. Reducción del indeterminismo en las preguntas de test

Con el objetivo de obtener variabilidad en sus respuestas, ChatGPT está diseñado para introducir un grado de indeterminismo a la hora de responder. Esto significa que, dada una misma pregunta, ChatGPT puede proporcionar respuestas diferentes cada vez que se le realice. Esta característica, por tanto, debe ser considerada a la hora de analizar los resultados obtenidos por ChatGPT derivados de experimentos como el que se propone en este trabajo. Por esta razón, se analiza el efecto de este indeterminismo en las repuestas generadas para las preguntas de test.

En el caso de las preguntas de test, este efecto no determinista puede ser fácilmente analizado; ya que, el espacio de las respuestas está limitado a las cuatro posibles respuestas que se proporcionan en el examen. Aprovechando esta característica de las preguntas de test, hemos añadido un paso adicional en la metodología que consiste en realizar cien veces cada una de las preguntas. De esta manera, podemos obtener la distribución de las diferentes respuestas proporcionadas por ChatGPT para una misma pregunta. El análisis de esta distribución nos permite, en muchos casos, obtener una respuesta predominante que podemos considerar como la definitiva para esa pregunta.

Para la implementación de este proceso se ha usado la API de ChatGPT[6]. Mediante un script en Python, se ha realizado cada pregunta de test cien veces, usando una conversación nueva en cada iteración. Además, se le ha indicado a ChatGPT que responda solo con la letra de la respuesta, sin incluir más información. Finalmente, por cada pregunta, se ha contabilizado el número de veces que ha proporcionado cada una de las cuatro respuestas posibles para su análisis posterior.

III. Resultados

A. Respuestas a las preguntas de investigación

A continuación se analizan los resultados obtenidos organizados en base a las preguntas de investigación.

¿Es capaz ChatGPT de superar los exámenes? Sin un entrenamiento específico en los contenidos de las asignaturas evaluadas, ChatGPT ha sido capaz de superar 8 de 15 exámenes. La última columna del Cuadro I muestra la calificación obtenida en cada examen. La calificación mínima para superar cada examen es 5, mientras que la máxima es 10. Los exámenes de aquellas asignaturas con un guion (-) como calificación no han sido corregidos por sus correspondientes responsables. Se muestra una sola calificación si hemos analizado un examen final, o dos calificaciones si hemos analizado dos exámenes parciales. Este resultado por sí solo nos sugiere que los actuales LLM ya tienen un impacto significativo y real en los métodos de evaluación de gran parte de las asignaturas evaluadas.

¿Acierta más preguntas de las que falla? Para obtener un dato de una granularidad más fina sobre el desempeño de ChatGPT en el examen, podemos analizar cuántas preguntas de examen ha acertado. Como muestra la Fig. 2, ChatGPT ha sido capaz de responder correctamente al 56 % de las preguntas que se le han formulado. En concreto, de un total de 230 preguntas ha respondido correctamente a 129. Si la evaluación se tratase de un solo examen y todas las preguntas tuviesen el mismo valor, ChatGPT habría superado la evaluación de todos los conocimientos considerados en el experimento. Sirva esta hipótesis para reflejar el impacto que tiene ChatGPT en los exámenes como método de evaluación.

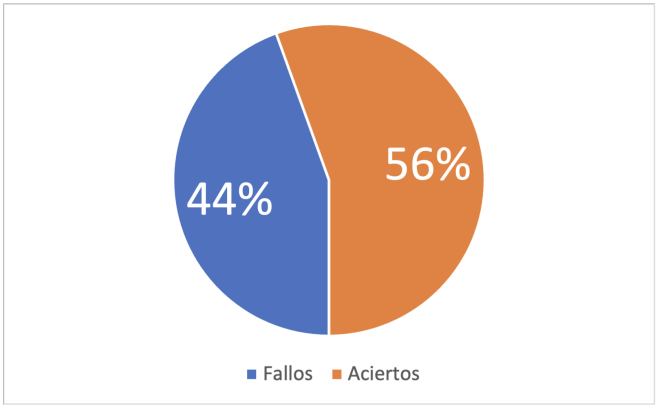

Fig. 2. Preguntas falladas y acertadas.

¿Influye el tipo de pregunta de examen? La Fig. 3 muestra tanto las cantidades de cada categoría de pregunta como el número de aciertos y fallos.

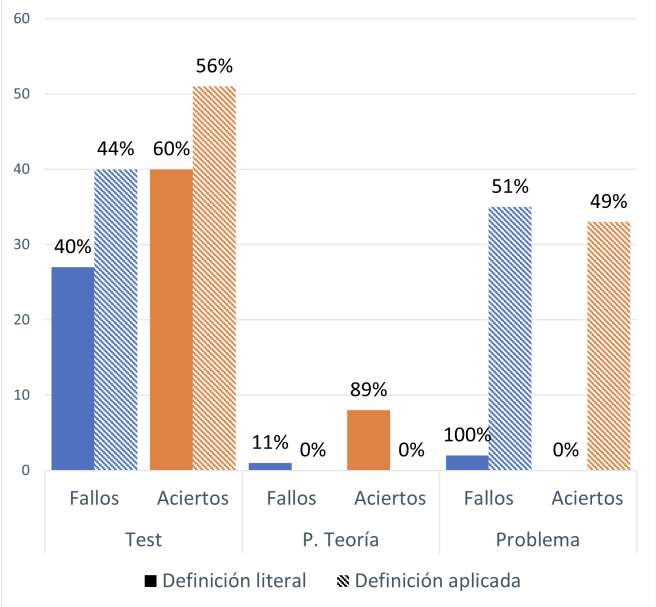

Fig. 3. Fallos y aciertos por tipo de pregunta.

[6] https://platform.openai.com/docs/api-reference

Curso	Nombre	Siglas	Tipo	Nota
1	Estructuras de datos y de la información	EDI	Test	4,58
1	Introducción a la programación	IP	Test & Problemas	3,90 + 7,88
2	Análisis y diseño de algoritmos	ADA	Test & Problemas	5,88 + 5
2	Bases de datos	BD	Test & Problemas	2,32
2	Desarrollo de programas	DP	Preguntas	7
2	Inteligencia artificial y sistemas inteligentes	IASI	Problemas	1,25
2	Programación concurrente y distribuida	PCD	Test & Problemas	5,05
3	Diseño y administración de bases de datos	DADB	Test & Problemas	3,39 + 2,19
3	Diseño e interacción en sistemas de información	DISI	Test	2,88
3	Diseño y modelado de sistemas software	DMSS	Preguntas & Problemas	6,45
3	Ingeniería de requisitos	IR	Preguntas	-
3	Programación en Internet	PI	Preguntas	7,5
3	Teoría de lenguajes	TL	Test & Problemas	2
4	Arquitecturas software en entornos empresariales	ASEE	Preguntas	5,25
4	Gestión de proyectos software	GPS	Preguntas	-

Cuadro I. Lista de asignaturas.

Como se puede observar, la gran mayoría de preguntas son de tipo test (155), seguidas por los problemas (70), siendo mucho menor el número de preguntas de teoría (9). Sin embargo, observando el porcentaje de aciertos por cada tipo de pregunta, podemos ver que el más alto se da en las preguntas de teoría, un 89 %, seguido por las preguntas de test, un 58 %, y finalmente, los problemas que presentan un ratio de aciertos por debajo del 50 %. Finalmente, si consideramos la dimensión de tipo de conocimiento, lo primero que podemos ver es que todas las preguntas de teoría se han clasificado como definición literal, mientras que los problemas pertenecen, en su mayoría, a la categoría de definición aplicada. En el caso de las preguntas de test, aparecen preguntas tanto en una como en otra categoría de esa dimensión y parece que ChatGPT presenta un ratio de aciertos bastante cercano en ambas categorías, un 60 % frente a un 56 %. Como ilustra la Fig. 4, considerando solo la dimensión de tipo de conocimiento, el porcentaje de aciertos en preguntas de definición literal es superior al de definición aplicada, un 62 % frente a un 53 %.

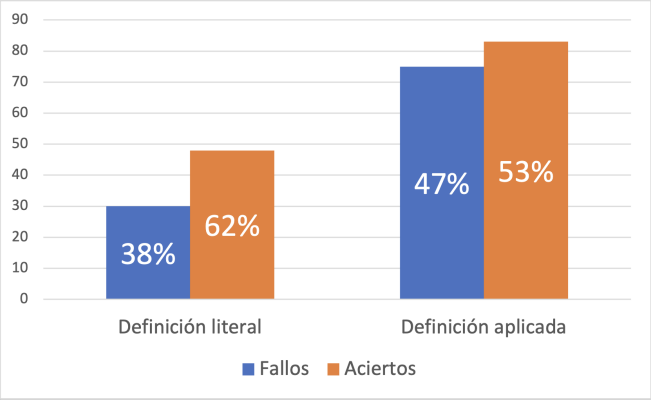

Fig. 4. Fallos y aciertos por tipo de conocimiento

Como primeras conclusiones, podemos comentar que estos resultados parecen coherentes con las capacidades de los LLM en el momento de realizar este estudio, que presentan un conocimiento formal del lenguaje mucho mayor que su conocimiento funcional [19]. Por lo tanto, es normal que sean capaces de repetir una definición literal pero tengan más problemas a la hora de aplicar la definición de un concepto a un ejemplo concreto. No obstante, nos ha sorprendido que el ratio de acierto no sea mayor en preguntas de test dentro de la categoría definición literal. Este resultado puede tener múltiples explicaciones como, por ejemplo, que el enunciado sea ligeramente ambiguo o las respuestas puedan solaparse. Sin embargo, no hemos detectado un número de casos significativo de preguntas que contengan este tipo de problemas. Tras un análisis más profundo, creemos que los fallos en las preguntas de test pueden estar asociados al efecto *mispriming* [20], que básicamente consiste en usar algún tipo de distractor en una pregunta. En este caso, las propias respuestas de los tests pueden estar actuando como distractores y llevar a ChatGPT a devolver una respuesta errónea. Aunque no tenemos una evidencia incontestable de este fenómeno, hemos probado en varias preguntas de test a proporcionarle solamente el enunciado sin incluir las respuestas y, en ese caso, ChatGPT ha respondido adecuadamente.

¿Influye el tipo de aplicación del conocimiento?

Como aparece en la Fig. 5, los tipos de aplicación más comunes son claramente el análisis de código, la aplicación de un algoritmo o método y la generación de código, en ese orden; mientras que los tipos de semántica operacional, cálculo matemático y análisis de diagramas se dan entre 10 y 20 preguntas.

En cuanto al ratio de aciertos, se obtienen valores de acierto bastante altos en análisis (62 %) y generación (70 %) de lenguajes de programación, mientras que en la aplicación de un algoritmo o método se producen más fallos que aciertos (46 %). Es reseñable también el elevado ratio de fallos que se producen en las categorías de cálculo matemático y análisis de diagramas.

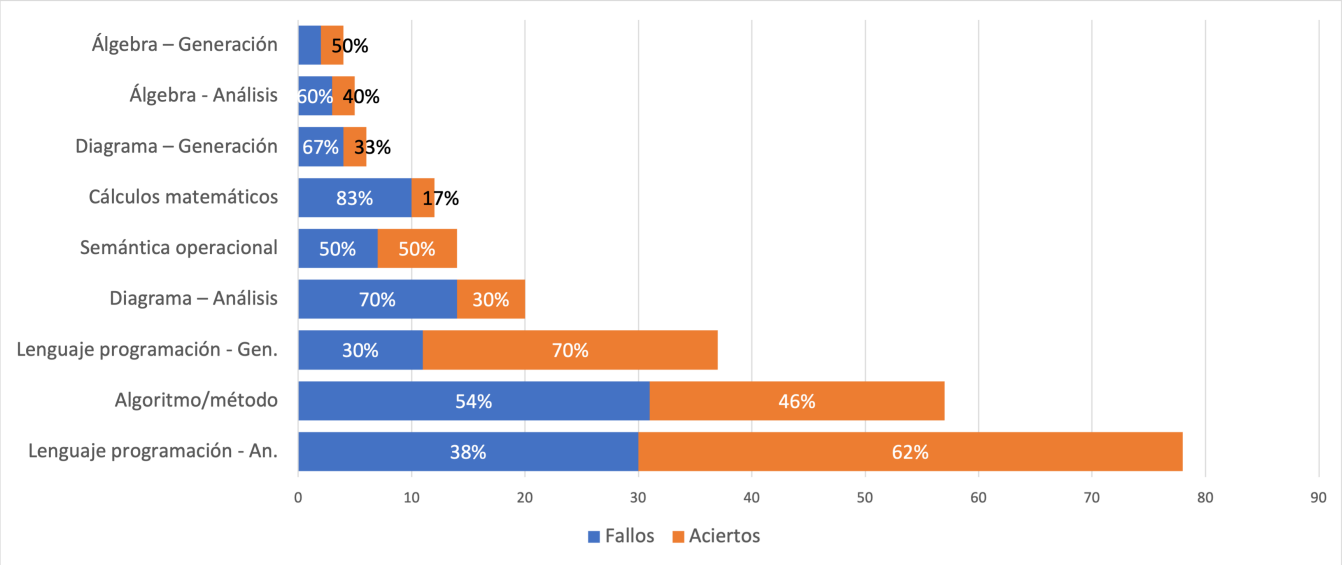

Figura 5. Fallos y aciertos por tipo de aplicación.

Como se esperaba, ChatGPT presenta un buen desempeño en preguntas de aplicación que requieren de competencias formales lingüísticas de algún lenguaje, como sucede en el caso de gran parte de las preguntas de análisis y generación de lenguajes de programación. Por otra parte, muestra un desempeño mucho más pobre en preguntas que requieren de competencias funcionales del lenguaje, como es el cálculo matemático o la aplicación de un algoritmo o método de múltiples pasos. Finalmente, en el caso relacionado con el análisis de diagramas, no se puede sacar ninguna conclusión válida, ya que tampoco podemos asegurar que la transcripción realizada de estos diagramas haya sido la más adecuada.

¿Cómo ha sido el desempeño de ChatGPT por asignatura?

Aparte de la calificación final obtenida, cabe hacer un análisis más pormenorizado de su desempeño por asignatura desde el punto de visto del tipo de preguntas de examen utilizadas, según la clasificación propuesta. Por cuestiones de brevedad, en este trabajo, solo se exponen con detalle los resultados de una asignatura. En concreto, se ha elegido la asignatura IP, en la que se han analizado dos exámenes parciales con resultados dispares. Se trata de la única asignatura en la que ChatGPT no ha presentado un desempeño homogéneo. El primer parcial no lo supera, mientras que en el segundo obtiene un notable alto. Cada parcial está compuesto por un test de 8 preguntas y un problema.

El Cuadro II muestra los resultados con mayor detalle para las preguntas de test. Como se puede ver, ChatGPT acierta todas las preguntas del segundo parcial, mientras que en el primero tiene el mismo número de aciertos y fallos. Atendiendo al tipo de conocimiento, ambos parciales contienen solo preguntas de definición aplicada. En cuanto a los distintos tipos de aplicación, todas, excepto una, requieren analizar algún fragmento de código. Por lo tanto, las principales diferencias entre ambos parciales se dan en los tipos de aplicación: semántica operacional y algoritmo/método. Solo el segundo parcial contiene tres preguntas de aplicación de algoritmo y todas son respondidas correctamente; por lo que, dado los resultados y el reducido número de preguntas de este tipo, tampoco parece que pueda explicar la diferencia de desempeño. Sin embargo, en el caso de la semántica operacional, sí tenemos un mayor número de preguntas y con resultados distintos. En concreto, centrándonos en el primer parcial, encontramos seis preguntas de este tipo de las cuales ChatGPT acierta dos y falla cuatro. Aunque el número de preguntas a analizar es reducido, podríamos concluir, en este caso, que ChatGPT presenta más dificultades cuando se le pregunta por el resultado de la ejecución de un fragmento de código.

P.	**Tipo**	1	2	3	4	5	6	7	8
1	LPA	✓	✓	✓	✓	✓	✓	✓	✓
1	SO	✓	✓	✓		✓		✓	✓
1	F/A	F	F	A	A	A	A	F	F
2	LPA	✓	✓	✓		✓	✓	✓	✓
2	SO						✓		
2	AM		✓		✓			✓	
2	F/A	A	A	A	A	A	A	A	A

Cuadro II. Resultados de IP.

B. *Análisis del indeterminismo en las preguntas de test*

En este análisis se han considerado 147 preguntas de test de las 155 disponibles, porque se han descartado todas aquellas que eran interdependientes y conllevaban una conversación larga con ChatGPT. La Fig. 6 muestra la variación de aciertos y fallos entre el estudio previo y el nuevo tomando como referencia la respuesta predominante. Como se puede apreciar, en términos totales, la reducción del efecto no determinista de ChatGPT ha supuesto obtener un incremento en el número de fallos. En concreto, se ha pasado de 61 fallos a 71, de un 41,5 % a un 48,3 % lo que supone un incremento de casi un 7 %. La última fila de esa figura muestra el número de preguntas que han cambiado de acierto a fallo o viceversa y, además, las que han cambiado de una respuesta incorrecta a otra (de fallo a fallo). Como se puede observar, 89 repuestas de 147 no

han cambiado, el 60,5 % por lo tanto coincide la respuesta original con la predominante obtenida. Sin embargo, el 39,5 % restante (58 respuestas) ha cambiado de la siguiente manera: 17 de fallo a acierto, 30 de acierto a fallo y 11 de fallo a fallo.

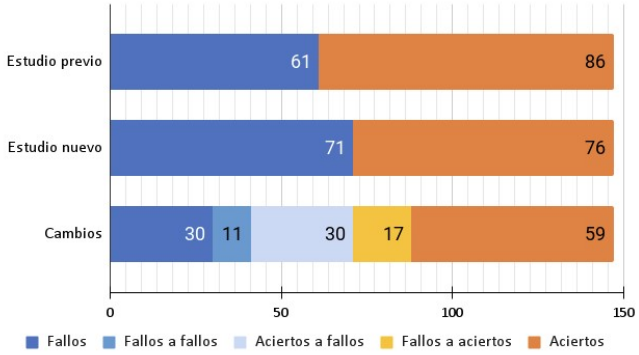

Fig. 5. Variación de aciertos y fallos.

En la Fig. 7 se muestra un histograma con la distribución de todos los aciertos y fallos según el valor porcentual de la respuesta predominante, agrupados en particiones de 10 en 10. Como se puede apreciar, las respuestas correctas (aciertos) aparecen con mayor frecuencia en las particiones más altas (80-100 % de valor de la respuesta predominante). Mientras que en el caso de las respuestas incorrectas (fallos) se distribuyen de una manera bastante uniforme por todas las particiones, presentando un mayor número en las particiones intermedias (50-80 %). Si bien, como se puede observar, se dan también casos extremos en los que la respuesta presenta un 100 % de predominancia pero es incorrecta. Y, por el otro lado, respuestas con baja predominancia, por debajo del 50 %, son correctas.

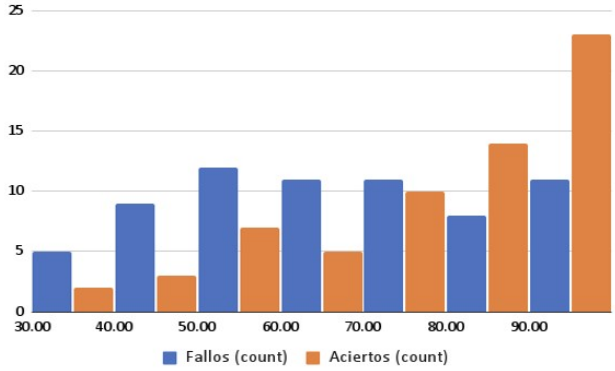

Fig. 6. Histograma de los porcentajes de la respuesta predominante categorizados por aciertos y fallos.

Por su parte, la Fig. 8 muestra la frecuencia de la respuesta predominante categorizada por el tipo de cambio para las 58 preguntas que presentan una nueva respuesta.

Como se puede apreciar, la mayoría de los cambios de fallo a acierto presentan una frecuencia claramente superior, por encima del 55 %, a los cambios de fallo a fallo (por debajo del 50 %). Esta situación puede indicar, por un lado, que en los cambios de fallo a acierto la respuesta del estudio original (fallo) no coincide con la más frecuente y, por otro lado, que en los cambios de fallo a fallo existen varias respuestas con una frecuencia similar, con lo que se da una mayor probabilidad de fallo. Finalmente, los cambios de acierto a fallo son los más numerosos y aparecen mucho más distribuidos en la figura. Por lo tanto, contienen tanto casos cuya respuesta correcta posee una frecuencia cercana a la predominante, pero menor, como casos cuya respuesta correcta presenta una frecuencia muy baja o incluso nula.

Aunque sería interesante realizar un análisis detallado de las preguntas cuya respuesta ha cambiado, esta parte del estudio queda fuera de este trabajo por falta de espacio para su desarrollo. No obstante, uno de los resultados más interesantes consiste en la identificación de preguntas aparentemente sencillas que ChatGPT falla con mucha frecuencia sin una razón aparente o fácil de identificar. Para ejemplificar este caso, se ha seleccionado una pregunta de test de la asignatura Introducción a la programación (IP). En concreto, la primera pregunta del primer parcial de IP, presentada en la Fig. 9, que entendemos constituye un ejemplo muy clarificador de cómo puede afectar el uso de distractores a ChatGPT.

Como podrá deducir cualquier lector familiarizado con la programación, la respuesta correcta es la *c*, atendiendo a la salida esperada de la ejecución del programa. Mientras que del primer parámetro (`a`) se obtiene una copia dentro del módulo `intercambiar`, del segundo (`b`) se obtiene una referencia. Por lo tanto, tras la invocación del módulo `intercambiar` sólo el valor del parámetro `b` resulta modificado, obteniéndose el resultado que se busca.

En las pruebas realizadas para el estudio original de este trabajo, ChatGPT respondió incorrectamente a esta pregunta, devolviendo la respuesta *d*. En un primer momento, este resultado se atribuyó a la naturaleza de la pregunta (definición aplicada de tipo semántica operacional), que requiere que la herramienta sea capaz de simular la ejecución del programa. Sin embargo, tras repetidos intentos, en una ocasión ChatGPT respondió correctamente a la pregunta, seleccionando la respuesta *c* como correcta. Posteriores pruebas, no obstante, volvieron a dar la respuesta *d* como correcta, dejando de manifiesto que ChatGPT no es un sistema con un comportamiento determinista.

La columna izquierda de la Fig. 10 muestra la distribución de las respuestas proporcionadas por ChatGPT a esa pregunta. La respuesta *d* es la más frecuente, prácticamente dos terceras partes de las veces, con un 63 %, seguida de la respuesta *c*, con un 24 %. El porcentaje de veces que ChatGPT selecciona la opción correcta no es despreciable, acertando casi una cuarta parte de las veces. Surge la pregunta de a qué puede deberse que la balanza se decante por la opción incorrecta más veces que por la correcta. ¿Podría deberse a la forma en la que la pregunta está formulada? A fin de cuentas, los LLM escogen la siguiente palabra de su respuesta de forma probabilística.

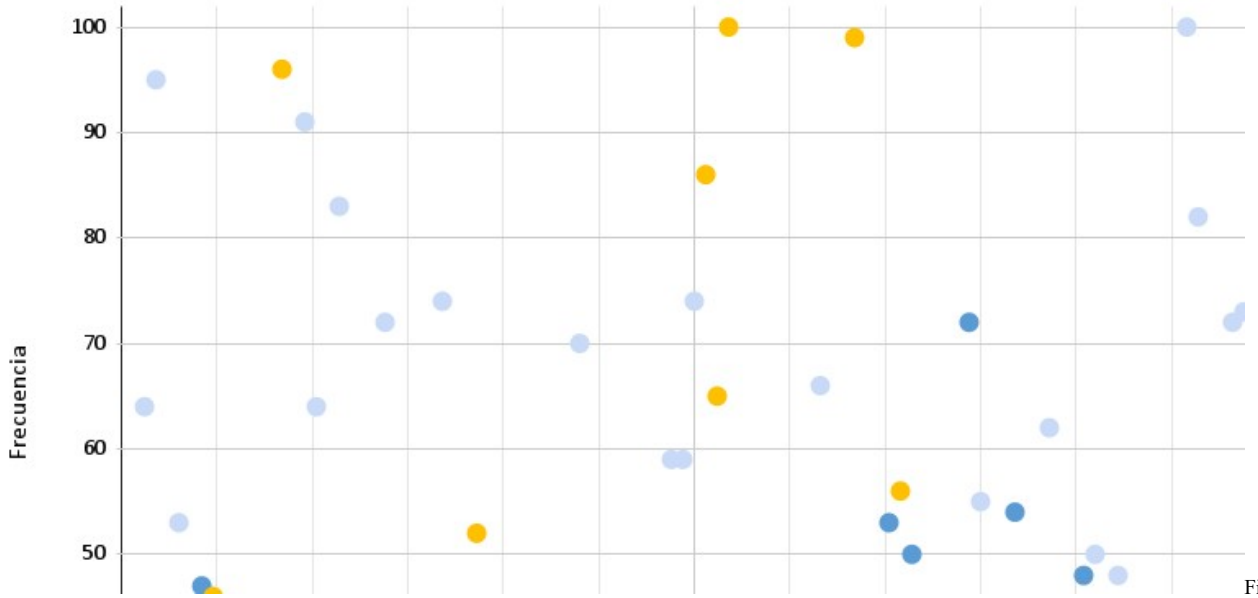

Figura 8. Frecuencia de las respuestas predominantes categorizadas por tipo de cambio.

Pregunta 1

Tenemos el siguiente módulo para intercambiar los valores de dos variables, pero no sabemos cómo son los parámetros:

```
void intercambiar ???????????{
        int aux;
        aux = x;
        x   = y;
        y   = aux;
}
```

Se ejecuta el siguiente algoritmo:

```
int main(){
        int a, b;
        a = 1;
        b = 2;
        intercambiar (a,b);
        cout << a << " " << b;
}
```

En la pantalla se escribe: **1 1**

¿Cómo era la cabecera de la función *intercambiar*?

a) `void intercambiar (int x, int y)`
b) `void intercambiar (int &x, int y)`
c) `void intercambiar (int x, int &y)`
d) `void intercambiar (int &x, int &y)`

Fig. 7. Primera pregunta del primer parcial de IP.

Tras analizarlo concienzudamente, nos percatamos de que, aunque el módulo se denominaba intercambiar, realmente el programa no realizaba un intercambio como tal del valor de ambos parámetros. En ese sentido, la respuesta d es la que debería seleccionarse si se buscase una cabecera que intercambiase los valores pasados como parámetros. ¿Qué ocurriría si se reformulase la pregunta, sustituyendo intercambiar por un nombre sin significado como foo? La Fig. 10 muestra la distribución de las respuestas proporcionadas por ChatGPT a la misma pregunta con solo realizar dicha sustitución. Como puede observarse, con este cambio, la respuesta c pasa a ser la más frecuente, superando el porcentaje obtenido originalmente por d, con un 69 %. Ninguna de las otras dos respuestas llega al 20 %.

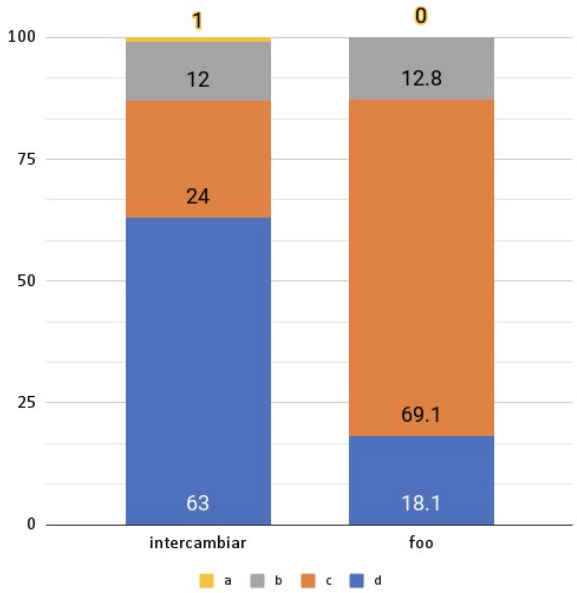

Fig. 8. Comparación de la distribución de respuestas en pregunta con distractor.

Como se demuestra en [20], es posible confundir con facilidad a un modelo de lenguaje preentrenado (*Pretrained Language Model*, PLM) como ChatGPT mediante el uso de *distractores* (*mispriming* en inglés), que consiste en añadir términos al enunciado de una pregunta para dirigir al modelo hacia una respuesta equivocada. En conclusión, aunque ChatGPT no responde la totalidad de las veces correctamente, el resultado obtenido en este ejemplo deja claro el papel del nombre del módulo como distractor en este caso.

IV. Recomendaciones

La variedad en el tipo de preguntas de un examen permite evaluar distintos niveles de conocimiento del alumnado en una determinada materia. En este trabajo, se realiza una serie de recomendaciones acerca del diseño de las preguntas para que se pueda valorar su inclusión o no en un determinado examen. Por supuesto, el profesorado deberá valorar estas recomendaciones en cada contexto y momento de evaluación. Por ejemplo, el riesgo del uso ilícito de ChatGPT en un examen por parte del alumnado puede ser diferente en modalidades presenciales o virtuales de enseñanza, o bien en pruebas de autoevaluación comparado con convocatorias oficiales de examen.

Como primera recomendación, se propone evitar, en la medida de lo posible, preguntas de definición literal, principalmente en preguntas de teoría. ChatGPT es capaz de responder preguntas complejas de este tipo, incluso razonando la respuesta y poniendo ejemplos de aplicación. En nuestro experimento, parece haber tenido más problemas en las preguntas de test, quizás por el hecho de que las propias respuestas pueden llegar a confundirlo, posiblemente por el efecto *mispriming*.

Como segunda recomendación, dentro de las preguntas de aplicación, se propone también reducir o evitar preguntas que solo supongan el análisis de código o la generación de código. ChatGPT es capaz de realizar un análisis estático del código bastante exhaustivo y también es capaz de generar fragmentos de código para resolver problemas bien acotados, como suelen ser los que se utilizan en preguntas de test en un examen. Al fin y al cabo, este tipo de competencia para un LLM no parece muy distinta a la que necesita para responder a preguntas de definición literal. Por lo tanto, es recomendable mezclar este tipo de aplicación con otros como la necesidad de conocer un algoritmo o método de múltiples pasos, o bien comprender la semántica operacional del código, o incluso la realización de cálculos matemáticos complejos. Por ejemplo, si se pide que genere el código de poda Alfa-Beta de un árbol de jugadas generado con el algoritmo minimax, también deberían realizarse preguntas sobre cuál sería el resultado al aplicarlo sobre un árbol concreto.

Como tercera recomendación, se propone el planteamiento de problemas que supongan la aplicación paso a paso de un algoritmo o un método de múltiples pasos. En este tipo de aplicación del conocimiento, ChatGPT ha tenido un desempeño bastante pobre, como puede verse por los resultados obtenidos en IASI, cuyos problemas son prácticamente todos de este tipo.

La cuarta recomendación consiste en incorporar preguntas con cálculos matemáticos complejos en los que, por ejemplo, sea necesario seguir correctamente un método determinado. Por ejemplo, en la asignatura DABD, una de sus respuestas contiene el siguiente error: ``*La tabla VUELO tendrá un tamaño de registro de 34 bytes (4 + 3 + 3 + 4 + 8 + 70)* ".

En quinto lugar, simplemente apuntar que el uso de diagramas que representen instancias concretas de estructuras de datos o modelos sobre los que haya que aplicar los conceptos aprendidos en la asignatura siempre supone un inconveniente adicional para el uso de ChatGPT, dado que es necesario realizar una traducción previa a un formato textual que pueda entender.

Como recomendación final, hemos detectado que supone un esfuerzo añadido proporcionarle a ChatGPT aquellos exámenes que plantean problemas a partir de un supuesto práctico suficientemente elaborado y complejo, como es el caso de BD o DABD, algo que parece afectar más claramente a su desempeño.

Es importante recordar que la validez de estas recomendaciones debe comprobarse siempre en la última versión de ChatGPT disponible, dada su evolución constante. Por ejemplo, en la versión del 30 de enero de 2023 se incorporan mejoras en cálculo matemático.

V. Conclusiones

En este trabajo, estamos interesados en evaluar las capacidades de ChatGPT para la superación de exámenes de un grado de Ingeniería Informática. Con este objetivo, hemos diseñado y desarrollado un experimento para proporcionar a ChatGPT los exámenes de 15 asignaturas. Los resultados obtenidos nos permiten concluir que este tipo de tecnologías ya tienen un impacto evidente en los métodos de evaluación que se usan en la educación superior. Por lo tanto, es necesario obtener una evaluación sistemática de sus capacidades para poder adaptar los métodos de evaluación de manera conveniente en cada contexto y momento. En este sentido, este trabajo propone una serie de recomendaciones en cuanto al diseño de preguntas de examen para los grados de Ingeniería Informática. Además, dado el carácter no determinista de ChatGPT, para obtener una medida real de su desempeño es fundamental realizar un análisis de la distribución de las respuestas proporcionadas en el caso de preguntas de test, al menos, como se ilustra en este trabajo.

Como principales líneas futuras de trabajo, podemos apuntar las siguientes. Primera, el desarrollo de un marco de referencia para la evaluación de las capacidades de este tipo de tecnologías dentro de las competencias propias de la Ingeniería Informática, que nos permita obtener una medida clara y rápida del desempeño de nuevas herramientas o de sus evoluciones. Además, existen otros LLM comparables a GPT-3 de OpenAI como, por citar algunos, OPT de Meta [7], o BLOOM [8]. Segunda, la personalización de LLM con contenidos específicos de los grados de informática para su aplicación innovadora en contextos de educación superior. Podría resultar interesante explorar el concepto de *estupidez artificial*, introducido de forma implícita por Alan Turing [21] y explícita por Lars Liden [22]. Un modelo ajustado con datos incorrectos se utilizaría para generar ensayos. El alumno tendría que analizar el trabajo para encontrar las inconsistencias. En ambos casos, los LLM construidos liberarían al docente de trabajo repetitivo como construir cuestionarios ajustados al temario o solucionar dudas básicas. Tercera, la utilización de ChatGPT como sistema de depuración de preguntas de los exámenes para, por ejemplo, reducir el contexto implícito usado en una pregunta, evitar respuestas de tests ambiguas o con solapamiento, o también eliminar posibles distractores de los enunciados, co-

mo se apunta en el ejemplo de pregunta de test con distractor mostrado al final de la Sección III-B, ya que puedan entorpecer la correcta comprensión de los mismos. A partir de un análisis más detallado de las distribuciones de las respuestas y su comparación con los fallos más frecuentes cometidos por el alumnado, se podría profundizar en esta línea de trabajo.

REFERENCIAS

[1] E. e. Tovar Caro, "Actas de las III Jornadas de Enseñanza Universitaria de Informática, Jenui 1997," Madrid, junio 1997.

[2] J. Divasón, F. J. Martínez de Pisón, A. Romero, and E. Sáenz de Cabezón, "Modelos de inteligencia artificial para asesorar el proceso evaluador de trabajos informáticos complejos," in Actas de las XXVII Jornadas de Enseñanza Universitaria de Informática, Jenui 2021. Asociación de Enseñantes Universitarios de la Informática (AENUI), 2021.

[3] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," The bulletin of mathematical biophysics, vol. 5, no. 4, pp. 115–133, 1943.

[4] OpenAI, "ChatGPT: Optimizing Language Models for Dialogue," Nov. 2022, disponible en https://openai.com/blog/chatgpt/. [Online]. Available: https://openai.com/blog/chatgpt/

[5] A. Vaswani et al., "Attention is All you Need," Advances in neural information processing systems, vol. 30, 2017.

[6] T. B. Brown et al., "Language Models are Few-Shot Learners," Advances in Neural Information Processing Systems, vol. 33, 2020.

[7] S. Zhang et al., "OPT: Open pre-trained transformer language models," arXiv:2205.01068, 2022, disponible en arXiv: https://arxiv.org/abs/2205.01068.

[8] B. Workshop, "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model," arXiv:2211.05100, 2022, disponible en arXiv: https://arxiv.org/abs/2211.05100. [Online]. Available: https://arxiv.org/abs/2211.05100

[9] A. Gilson et al., "How does ChatGPT perform on the medical licensing exams? The implications of large language models for medical education and knowledge assessment," 2022, disponible en medRxiv: https://www.medrxiv.org/content/10.1101/2022.12.23.22283901v1.

[10] M. J. Bommarito and D. M. Katz, "GPT Takes the Bar Exam," 2022, disponible en SSRN: https://ssrn.com/abstract=4314839.

[11] S. Wollny, J. Schneider, D. Di Mitri, J. Weidlich, M. Rittberger, and H. Drachsler, "Are we there yet? - A systematic literature review on chatbots in education," Frontiers in Artificial Intelligence, vol. 4, 2021.

[12] X. Zhai, "ChatGPT user experience: Implications for education," 2022, disponible en SSRN: https://ssrn.com/abstract=4312418.

[13] M. M Alshater, "Exploring the role of artificial intelligence in enhancing academic performance: A case study of ChatGPT," 12 2022, disponible en SSRN: https://ssrn.com/abstract=4312358. [Online]. Available: https://ssrn.com/abstract=4312358

[14] T. Susnjak, "ChatGPT: The End of Online Exam Integrity?" arXiv:2212.09292, 2022, disponible en arXiv: https://arxiv.org/abs/2212.09292.

[15] P. A. C. Debby R. E. Cotton and J. R. Shipway, "Chatting and cheating: Ensuring academic integrity in the era of chatgpt," Innovations in Education and Teaching International, vol. 0, no. 0, pp. 1–12, 2023. [Online]. Available: https://doi.org/10.1080/14703297.2023.2190148

[16] Y. K. Dwivedi et al., "Opinion paper: "so what if chatgpt wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy," International Journal of Information Management, vol. 71, p. 102642, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0268401223000233

[17] F. J. García Peñalvo, F. Llorens-Largo, and J. Vidal, "La nueva realidad de la educación ante los avances de la inteligencia artificial generativa," RIED-Revista Iberoamericana de Educación a Distancia, vol. 27, no. 1, p. 9–39, ene. 2024. [Online]. Available: https://revistas.uned.es/index.php/ried/article/view/37716

[18] R. Rodriguez-Echeverria, J. D. Gutiérrez, J. M. Conejero, and A. E. Prieto, "Impacto de chatgpt en los métodos de evaluación de un grado de ingeniería informática," in Actas de las XXIV Jornadas de Enseñanza Universitaria de Informática, Jenui 2023. Asociación de Enseñantes Universitarios de la Informática (AENUI), 2023.

[19] K. Mahowald, A. A. Ivanova, I. A. Blank, N. Kanwisher, J. B. Tenenbaum, and E. Fedorenko, "Dissociating language and thought in large language models: A cognitive perspective," arXiv:2301.06627v1, 2023, disponible en arXiv: https://arxiv.org/abs/2301.06627.

[20] N. Kassner and H. Sch¨utze, "Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7811–7818.

[21] A. M. Turing, "I.—Computing Machinery and Intelligence," Mind, vol. LIX, no. 236, pp. 433–460, 10 1950. [Online]. Available: https://doi.org/10.1093/mind/LIX.236.433

[22] L. Lidén, "Artificial stupidity: The art of intentional mistakes," AI game programming wisdom, vol. 2, pp. 41–48, 2003.G. O. Young, "Synthetic structure of industrial plastics (Book style with paper title and editor)," in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.

Roberto Rodríguez-Echeverría es Profesor Titular de arquitectura de software en el Departamento de Ingeniería de Sistemas Informáticos y Telemáticos de la Universidad de Extremadura (UEx), España. Cuenta con una experiencia docente de más de 20 años en el ámbito de la ingeniería del software. Sus intereses de investigación incluyen la ingeniería de software, la educación de la ingeniería del software y la inteligencia artificial. Actualmente es director del Instituto de Tecnología Informática Aplicada.

Juan D. Gutiérrez es Profesor Ayudante Doctor en el Área de Lenguajes y Sistemas Informáticos del Departamento de Electrónica y Computación en el Campus de Lugo de la Universidad de Santiago de Compostela (USC) y tiene más de veinte años de experiencia en el mundo de la informática. Recientemente ha presentado su doctorado en el campo de los sistemas de posicionamiento en interiores (*Indoor Positioning Systems*, IPS) basados en luz LED visible. Actualmente su investigación está centrada en la aplicación de la inteligencia artificial a diferentes campos de conocimiento. Su formación incluye la programación en diferentes lenguajes, la administración de sistemas, el diseño de aplicaciones y las bases de datos e Internet.

José M. Conejero es Profesor Titular del Departamento de Ingeniería de Sistemas Informáticos y Telemáticos de la Universidad de Extremadura (España), donde ha impartido varias asignaturas relacionadas con la Programación y la Ingeniería del Software. Se doctoró en Informática por la Universidad de Extremadura en 2010. Es autor de más de 50 artículos de revistas y actas de congresos y también ha participado en diferentes revistas y congresos como miembro del comité de programa. Sus áreas de investigación incluyen la Ingeniería Web, el Big Data y el Desarrollo Dirigido por Modelos.

Álvaro E. Prieto es Profesor Titular de Lenguajes y Sistemas Informáticos en la Universidad de Extremadura (España). Es miembro del Grupo de Ingeniería del Software Quercus. Se licenció en Informática por la Universidad de Extremadura en 2000 y se doctoró en Informática en 2013. Sus intereses de investigación incluyen Linked Open Data, Predictive Analytics y Business Intelligence. Actualmente participa en varios proyectos de I+D+i.