Congreso In-Red 2024
UPV, 11 – 12 de julio de 2024
Doi: https://doi.org/10.4995.INRED2024.2024.18508

Uso del Modelo Generativo GPT-4 para la Elaboración de Cuestionarios de Evaluación en Asignaturas de Grado

Use of the GPT-4 Generative Model for Developing Assessment Questionnaires in Undergraduate Courses

Roberto Fernandez-Moran[1], Miguel-Ángel Fernández-Torres[1], Julia Amorós-López[1], Ana B. Ruescas[1], Valero Laparra[1], Maria Piles[1], Jordi Muñoz-Marí[1], Veronica Nieves[1], Jose E. Adsuara[1], Álvaro Moreno-Martínez[1] y Luis Gómez-Chova[1]

[1] Universitat de València, roberto.fernandez@uv.es

How to cite: Fernandez-Moran, M.; Fernández-Torres, M.A.; Amorós-López, J.; Ruescas, A.B.; Laparra, V.; Piles, M.; Muñoz-Marí, J.; Nieves, V.; Adsuara, J.E.; Moreno-Martínez, Á y Gómez-Chova, L. (2024). Uso del Modelo Generativo GPT-4 para la Elaboración de Cuestionarios de Evaluación en Asignaturas de Grado. En libro de actas: *X Congreso de Innovación Educativa y Docencia en Red.* Valencia, 11 – 12 de julio de 2024.
Doi: https://doi.org/10.4995/INRED2024.2024.18508

Abstract

This study examines the use of the generative model GPT-4 to generate multiple-choice questions in undergraduate subjects, by reading material in PDF format and other educational resources distributed to students in class. Multiple-choice questions were automatically generated for various subjects, and their suitability was evaluated quantitatively and qualitatively. The results show that GPT-4 provides a quick and flexible way to generate self-assessment quizzes and exams, which streamlines the assessment process and student learning. However, the relevance of the generated content depends largely on the information available on the internet, as well as the materials used by the teacher to train the model, with human supervision always necessary.

***Keywords**: AI, generative models, GPT-4, evaluation, automated questionnaires.*

Resumen

Este estudio analiza el uso del modelo generativo GPT-4 para generar preguntas tipo test en asignaturas de grado, mediante la lectura de material en PDF y otros recursos educativos que se distribuyen al alumnado en clase. Para diversas asignaturas, se generaron preguntas tipo test de forma automática y se evaluó de forma cuantitativa y cualitativa la idoneidad de las mismas. Los resultados muestran que GPT-4 ofrece una forma rápida y flexible d e g enerar c uestionarios a utoevaluables y e xámenes, l o cual agiliza el proceso de evaluación y el aprendizaje de los estudiantes. Sin embargo, la pertinencia del contenido generado depende en gran medida de la información presente en la red, así como de los materiales con los que el docente alimenta al modelo, siendo en todo caso necesaria la supervisión humana.

***Keywords**: AI, modelos generativos, GPT-4, evaluación, cuestionarios automáticos.*

*Este trabajo ha sido realizado en el marco del proyecto docente UV-SFPIE_PIEE-2737016, al que pertenecen todos los autores.

1 Introducción

El desarrollo de métodos para la enseñanza automática ha sido desde siempre un objetivo clave. Ya en los años 50, el psicólogo e inventor B.F. Skinner propuso unas máquinas automáticas para facilitar la enseñanza (Skinner, 1958). Skinner argumentaba que los métodos automáticos son clave para poder aplicar de forma eficiente el condicionamiento operante a la enseñanza, ya que uno de los requisitos para su correcto funcionamiento es proporcionar un refuerzo (o evaluación) lo más rápido posible al alumnado. En este sentido, la velocidad de respuesta que puede ofrecer un método automático frente a un profesor es claramente superior, dado que en general cada profesor ha de encargarse de más de 20 alumnos y esto dificulta la atención personalizada. Aunque la eficiencia para la enseñanza de las máquinas automáticas fue probada por Skinner, éstas eran caras de construir y el diseño de cada uno de los ejercicios era arduo y costoso.

En la era digital actual, la integración de tecnologías de inteligencia artificial (IA) en la educación está transformando la forma en que se enseñan y evalúan los conocimientos. Los educadores debemos reconocer y entender el *ecosistema tecnológico de aprendizaje* (Wilkinson, 2002) de nuestros alumnos, comprendiendo que lo que ocurre en el aula es sólo una pieza de un conjunto más amplio de influencias que contribuyen al progreso. El surgimiento de tecnologías de IA generativa, como GPT-4 (Baidoo-Anu y Owusu Ansah, 2023), está transformando radicalmente diversos aspectos de nuestra sociedad, desde la comunicación hasta la creación de contenido e información. Este rápido avance tecnológico ha generado cierta preocupación inicial en las instituciones educativas. Como bien indican Ribera y Díaz Montesdeoca, 2024 señalando el trabajo de Mike Sharples de la Open University UK, si bien las IA generativas representan un desafío disruptivo para la educación, también ofrecen grandes oportunidades para apoyar el aprendizaje, especialmente debido a su alta capacidad de comunicar de forma casi idéntica a los humanos mediante el uso del lenguaje natural.

La integración de los Modelos de Lenguaje de Gran Tamaño (LLM, por sus siglas en inglés, "Large Language Models") y la IA generativa, como GPT-4, en el ámbito educativo, abre nuevos caminos en la enseñanza y el aprendizaje. Los LLM se especializan en entender, generar y manipular lenguaje humano a gran escala. La inteligencia artificial generativa permite crear contenido multimedia nuevo y único a partir de patrones aprendidos en datos masivos. GPT-4, en particular, es un ejemplo avanzado de estas tecnologías, capaz de realizar tareas complejas de procesamiento del lenguaje natural con una comprensión contextual profunda. GPT-4, como IA generativa, no solo interpreta texto y contenido multimedia sino que también lo genera, adaptándose a una amplia gama de estilos y formatos según las necesidades.

En este estudio, hemos utilizado GPT-4 para leer y analizar documentos que típicamente se distribuyen al alumnado en clase, incluyendo manuales teóricos, prácticos, diapositivas, problemas y guiones de laboratorio, así como códigos de programación. A partir de la lectura de dichos documentos, se ha utilizado GPT-4 para la generación de cuestionarios automáticos autoevaluables. Para el acceso a GPT-4, se ha usado tanto la interfaz gráfica de GPT-4 como la API de OpenAI. La API permite el acceso a GPT4 utilizando el lenguaje de programación Python, lo que ofrece posibilidades ilimitadas para personalizar y optimizar aplicaciones educativas.

2 Objetivos

El objetivo principal de este estudio es investigar cómo un LLM basado en IA generativa, en concreto GPT-4, puede ser utilizado para generar preguntas tipo test, elaborando cuestionarios para un conjunto diverso de asignaturas de grado. Los objetivos específicos son los siguientes:

1. Evaluar la efectividad de GTP-4 en la generación de preguntas relevantes y precisas acerca de la temática aprendida en clase, tanto teórica como práctica.
2. Analizar la viabilidad de utilizar material en formato PDF y otros como contexto para la generación de dichas preguntas.
3. Evaluar la calidad de las preguntas generadas automáticamente y analizar su contribución en la mejora del proceso de aprendizaje y evaluación del alumnado.

> *Se desea principalmente mostrar el potencial de los modelos generativos de inteligencia artificial para mejorar la evaluación docente a través de la generación automática de preguntas de evaluación continua y exámenes.*

3 Desarrollo de la innovación

GPT-4, un modelo de lenguaje generativo basado en IA desarrollado por OpenAI, ha demostrado desde que se lanzó en marzo de 2023 una notable capacidad para comprender y generar texto coherente en una variedad de contextos. Este modelo mejora a sus predecesores, GPT-3 y GPT-3.5, lanzados respectivamente en mayo de 2020 y noviembre de 2022, siendo este último el que alcanzó gran popularidad a nivel mundial con el lanzamiento de la aplicación web ChatGPT. Esta capacidad puede ser aprovechada en el contexto educativo para generar preguntas tipo test que aborden conceptos clave dentro de diversas disciplinas académicas. La idea de partida es generar multitud de preguntas de manera automática, delimitando los resultados en base a criterios sencillos, a partir del material de aprendizaje distribuido en clase, ya sea de las presentaciones que se muestran o de las lecturas complementarias que se ponen a disposición de los estudiantes. A partir de las preguntas generadas, y tras la revisión del docente, pueden construirse multitud de cuestionarios de autoevaluación que se integran en otras herramientas de clase como *Moodle* de una manera casi automática.

3.1 Uso de GPT-4 en las asignaturas

En la Tabla 1 se incluyen las asignaturas de grado impartidas en la Universitat de València (www.uv.es) para las que se han generado preguntas tipo test utilizando GPT-4: Señales y Sistemas (SyS) y Modelos Conexionistas (MC), del Grado en Ciencia de Datos; Programación Concurrente y Distribuida (PRO), Sistemas Integrados de Fabricación (SIF) y Teoría de Redes Eléctricas (TRE), del Grado de Ingeniería Electrónica Industrial; y Biogeografía (BIGE), del Grado en Geografía y Medio Ambiente. También se indica en la tabla el número de preguntas generadas de forma automática para cada asignatura.

Tabla 1: Asignaturas para las que se han generado preguntas tipo test utilizando GPT-4.

Titulación	Asignatura	Nº Preguntas
Grado en Ciencia de Datos	Señales y Sistemas (SyS)	239
Grado en Ciencia de Datos	Modelos Conexionistas (MC)	80
Grado de Ingeniería Electrónica Industrial	Programación Concurrente y Distribuida (PRO)	40
Grado de Ingeniería Electrónica Industrial	Sistemas Integrados de Fabricación (SIF)	150
Grado en Geografía y Medio Ambiente	Biogeografía (BIGE)	110
Grado de Ingeniería Electrónica Industrial	Teoría de Redes Eléctricas (TRE)	70

Para ello, se han considerando dos escenarios diferentes:

1. En el primero de ellos, *sin contexto*, se utiliza la línea de comandos o *prompt* de GPT-4 únicamente, sin proporcionarle información adicional acerca de la asignatura o tema para el que se quieren obtener preguntas. Simplemente se describe la temática sobre la que se deben de realizar las preguntas. Además, se solicita la respuesta en formato Aiken, ya que este formato es uno de los más sencillos aceptados para ser importado en el Aula Virtual de la plataforma *Moodle* de la Universidad de Valencia. En el *prompt* de GPT-4 se realiza la siguiente consulta:

   > *Dado el tema ≪ X ≫ genera 10 preguntas teóricas de dificultad medio-alta con 4 respuestas en formato Aiken.*

2. En el segundo escenario, *contextualizado*, se facilita el material relacionado con la asignatura o temas para los que se quieren generar entre 10-30 preguntas y se envía un *prompt* a GPT-4 mediante la API de OpenAI. En este caso en concreto se ha utilizado el material en formato PDF, realizando la siguiente consulta:

   > *En base al texto de los PDF introducidos, genera 10 preguntas teóricas de dificultad medio-alta con 4 respuestas en formato Aiken.*

   Cabe destacar que, aunque se optó por el uso de la API de OpenAI, la generación de preguntas tras la lectura de PDF usando el modelo GPT-4 es también posible a través de la interfaz gráfica ChatGPT 4, aunque este último ofrece menor flexibilidad en el formato de salida de los resultados, limitándose a producir texto sin un formato específico.

Ejemplo de pregunta (formato Aiken):

¿Qué es una época en el contexto de entrenamiento de un perceptrón?
A) Una única actualización de los pesos.
B) Una iteración sobre un conjunto de datos completo.
C) El tiempo que tarda en converger el algoritmo.
D) El número de muestras procesadas por el perceptrón.
ANSWER: B

3.2 Integración de material en PDF y otros recursos

En la generación de preguntas usando contexto se ha utilizado la API de OpenAI para leer los diferentes documentos de los temas de las asignaturas en formato PDF. Para uniformizar las interacciones con GPT-4, se creó una plantilla en el entorno de desarrollo *Jupyter Notebook* con código en el lenguaje de programación *Python* que permitía generar cuestionarios de forma automática y guardarlos como ficheros. Cada uno de los docentes que intervinieron en este estudio usó uno o varios PDF referentes a varios temas tratados en la asignatura que impartía, proporcionando al modelo un contexto para generar preguntas. La lectura de los archivos PDF se realizó empleando la librería de *Python pdfplumber* para extraer texto, datos e imágenes. A partir de los mencionados documentos cada profesor utilizó GPT-4 para generar las preguntas y cuatro posibles respuestas, especificando su correspondiente respuesta correcta. Posteriormente el profesorado revisó y corrigió las preguntas generadas, en caso de requerirse, para poder ser utilizadas en el aula. Las preguntas originalmente creadas por GPT-4 y posteriormente modificadas por los docentes, fueron comparadas usando métricas de similitud de contenido (distancia de Jaccard) y una evaluación semántica basada en temas latentes.

3.3 Evaluación de la calidad de las preguntas generadas

3.3.1 Evaluación cualitativa de las preguntas generadas

En la evaluación cualitativa de las preguntas generadas por GPT-4 realizada por los docentes de cada asignatura, se han considerado las siguientes categorías:

- Utilizables
- Correctas
- Modificadas
- Inadecuadas
- Repetidas
- Erróneas
- Difíciles
- Fuera de contexto
- Fáciles
- Demasiado numéricas

3.3.2 Evaluación semántica basada en temas latentes

En aprendizaje máquina, Latent Dirichlet Allocation (LDA) (Blei et al., 2003) es un modelo generativo que permite explicar conjuntos de observaciones, en este caso las palabras que constituyen las preguntas tipo test para una determinada asignatura, a partir de grupos no observados o temas latentes que determinan el motivo por el que algunas partes de estos datos (preguntas) son similares. Para la evaluación semántica de conjuntos de preguntas tipo test, si las observaciones son palabras en dichas preguntas, cada una de las preguntas se puede definir como una mezcla de un número pequeño de temas o categorías, de manera que la presencia de cada palabra en una pregunta es debida a uno de los temas a los que la pregunta pertenece.

El modelo LDA se apoya en representaciones de las preguntas tipo test basadas en un modelo bolsa de palabras (del inglés *bag-of-words*) Goldberg, 2022. Siguiendo este modelo, cada una de las preguntas a analizar se corresponde con un vector de características numéricas, atendiendo al siguiente procedimiento de vectorización:

1. *Eliminación o filtrado de palabras vacías (stop words).* Dada una pregunta, las palabras vacías, tales como "en", "algo.º "como", sirven para dar coherencia y naturalidad a su texto correspondiente. Sin embargo, estas palabras no son relevantes a la hora de analizar el contenido semántico de la misma, por lo que no se tendrán en cuenta a la hora de representarla.

2. *Tokenización.* Para cada palabra o token en una pregunta, asumiendo espacios en blanco y signos de puntuación como separadores de tokens, se asigna un identificador numérico, siendo este identificador único para cada palabra y, por tanto, el mismo en todas las preguntas en las que aparezca.

3. *Contabilización (normalizada)* del número de ocurrencias de cada token en cada pregunta.

4. *Ponderación*, con importancia decreciente, de los tokens que aparecen en la mayoría de las preguntas tipo test del conjunto a estudio. Es decir, si una palabra aparece en muchas de las preguntas del conjunto, su ponderación es baja, mientras que palabras menos frecuentes reciben una ponderación mayor.

Dadas todas las palabras disponibles en un conjunto de preguntas, cada pregunta se representa, por tanto, de acuerdo a la ocurrencia de dichas palabras en la misma. Siguiendo esta aproximación, el orden de las palabras en las preguntas no importa al modelo, es decir, las palabras representan la misma información independientemente del lugar que ocupan. Esta asunción es necesaria para facilitar la determinación de la probabilidad de pertenencia de cada una de las palabras a cada uno de los temas que representan el conjunto de preguntas tipo test. A pesar de que esto puede implicar algunas veces el trato de frases semánticamente diferentes como similares, funciona bien en general para el análisis de documentos de texto.

En la evaluación semántica que se presenta en la Sección 4.2, el modelo LDA asume cada conjunto de preguntas como un todo (corpus) y determina los temas a partir del mismo. Si las preguntas se hubieran comparado de forma individual, ciertos temas podrían no ser recogidos, los cuales son solamente identificables cuando se observa el corpus de preguntas completo. Los temas obtenidos por LDA, definidos por la probabilidad de pertenencia de cada una de las palabras en el corpus a los mismos, se pueden interpretar de la siguiente manera: las palabras que aparecen con menos frecuencia en las preguntas individuales, pero que son comunes a muchas preguntas diferentes, probablemente son indicativas de que existe un tema común entre las preguntas. Por tanto, a la hora de resumir un conjunto de preguntas como los que analizamos en la sección de resultados, la capacidad de los temas de resumir su contenido permite que la información más relevante sea incluida con una menor probabilidad de repetición.

3.3.3 Evaluación cuantitativa de las preguntas generadas

La distancia de Jaccard (Jaccard, 1901) se emplea para cuantificar la disimilitud entre dos conjuntos de datos. En el contexto de nuestro estudio, esta métrica se utiliza para comparar el texto de las preguntas generadas por el modelo GPT-4 con las correcciones realizadas por docentes en diversos temas. Inicialmente, se procesaron los textos para normalizarlos y facilitar la comparación. Este proceso incluye la conversión a minúsculas y la eliminación de puntuación. Posteriormente, se separó en *tokens* el texto, es decir, se dividió en palabras individuales, y se generó un conjunto de *tokens* únicos para cada texto.

La distancia de Jaccard se define como:

$$J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

donde A y B son los conjuntos de palabras únicas de los textos a comparar.

3.3.4 Encuestas al profesorado sobre el uso de LLMs en el aula

Finalmente, la realización de encuestas permite evaluar la utilidad de los modelos LLM en el aula. Para extraer conclusiones adicionales acerca de la innovación desarrollada para este trabajo, se ha utilizado la encuesta del Anexo A para recabar la opinión del profesorado involucrado en la actividad.

4 Resultados

En esta investigación se utiliza el modelo GPT-4 para generar preguntas tipo test en varias asignaturas de grado, abarcando los ámbitos de matemáticas, ciencias de la computación, ingeniería y ciencias sociales. Observamos que las preguntas generadas por GPT-4 pueden abarcar una amplia gama de temas y ser relevantes como material de estudio y evaluación. Además, al utilizar material de cada asignatura en PDF como contexto, se encontró una mayor coherencia y especificidad en las cuestiones planteadas.

4.1 Análisis cualitativo de las preguntas generadas por asignaturas

Para cada asignatura, los docentes evaluaron las preguntas generadas por GPT-4 y las clasificaron en las distintas categorías (ver figura 1). En el estudio se observó que GPT-4 no solo demostró una notable capacidad para generar preguntas de forma autónoma en diversas asignaturas de grado, sino también para utilizar material en PDF, como diapositivas de presentaciones, guiones de sesiones de laboratorio y referencias bibliográficas, como contexto para generar preguntas más relevantes y contextualizadas. Este enfoque permite que las preguntas generadas estén más alineadas y acotadas al contenido específico que los estudiantes están aprendiendo en cada asignatura.

1. SyS (Señales y Sistemas):
    - En esta asignatura se generaron cuestiones que abordaban aspectos teóricos de la asignatura. Estas cuestiones se mezclaban con cuestiones más prácticas, como problemas cuantitativos, que se desarrollaron por el docente, ya que los LLMs aún no son capaces de abordar con eficiencia cuestiones numéricas. GPT-4 generó preguntas que abordaban conceptos fundamentales como la transformada de Fourier y la convolución. En general

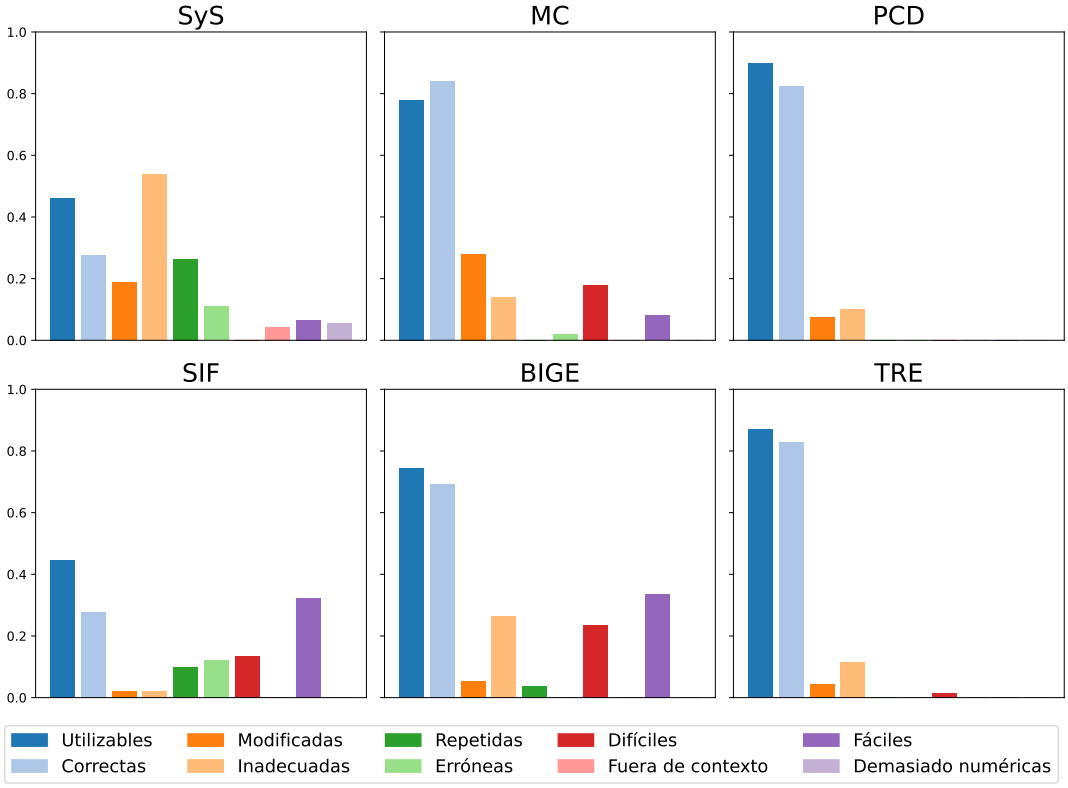


Fig. 1: Análisis cualitativo de las preguntas generadas por GPT-4 (uso contextualizado*) para cada una de las asignaturas consideradas en este estudio. SyS: Señales y Sistemas, MC: Modelos Conextionistas, PCD: Programación Concurrente y Distribuida, SIF: Sistemas Integrados de Fabricación, BIGE: Biogeografía, TRE: Teoría de Redes Eléctricas.*

estos conceptos están interiorizados dentro del LLM de modo que no había diferencia notable entre lo que generaba el LLM cuando no se le pasaba contexto y cuando se le pasaba contexto. En general el porcentaje de preguntas utilizables es suficientemente alto para que el sistema sea recomendable, es decir, aunque hay que repasar las preguntas generadas, el tiempo invertido es mucho menor que si se tuviesen que generar por el docente.

2. MC (Modelos Conexionistas):

   - En esta asignatura, se generaron cuestionarios automáticos a partir de los guiones de las prácticas de laboratorio, que incluían el enunciado de los ejercicios, una pequeña parte de marco teórico y algunos casos prácticos (códigos de programación en lenguaje Python). La materia abarca diferentes temas relacionados con los algoritmos de aprendizaje máquina aplicados en el campo de la inteligencia artificial. El hecho de que el modelo de GPT-4 se entrena con información *online* y que el campo de la inteligencia artificial está ampliamente documentado en la red por la naturaleza de las personas que

investigan y trabajan en él, pueden justificar el porcentaje significativamente alto de preguntas que resultaron útiles.

3. PRO (Programación Concurrente y Distribuida):

   - En esta asignatura se generaron cuestionarios automáticos para la parte teórica de la misma. GPT-4 generó preguntas que cubrían tanto conceptos básicos de programación orientada a objetos (clases, herencia, polimorfismo, abstracción, clases abstractas, interfaces, excepciones, etc.) como conceptos avanzados de programación concurrente y distribuida (sincronización de hilos con semáforos y monitores, la serialización de objetos y su envío mediante los protocolos UDP y TCP, etc.). Dado que se utilizó el mismo modelo para generar los cuestionarios que el que se había utilizado previamente para la generación de diapositivas teóricas en PDF (es decir, se disponía de contexto), la mayoría de las cuestiones generadas fueron clasificadas como utilizables.

4. SIF (Sistemas Integrados de Fabricación):

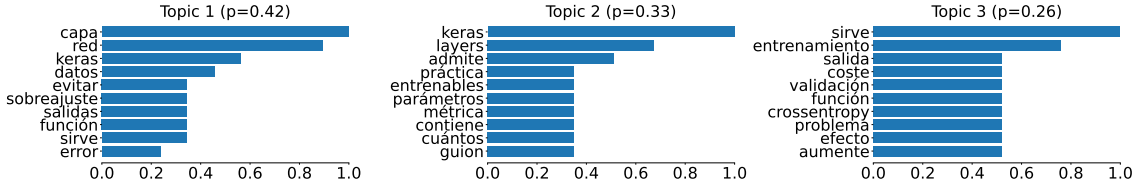
   - En esta asignatura se generaron cuestiones tipo test que serán utilizadas para los exámenes parciales y finales. Se trata de una asignatura con bastante contenido teórico, ya que se definen muchos conceptos de automatización y sistemas distribuidos aplicados a instalaciones industriales. También se definen muchos buses industriales, con protocolos muy específicos y para los que en la red abunda información básica, pero cuyos detalles son difíciles de encontrar o están desactualizados. Debido a esto y a pesar de utilizar contexto, bastantes preguntas eran o bien excesivamente simples o muy complejas y se basaban en datos de tablas de características muchas veces no demasiado relevantes. También se observó que al requerir un número elevado de preguntas (más de 30), se generaban contenidos repetitivos. Cabe destacar también que en este caso el objetivo era generar preguntas de examen, por lo que el grado de exigencia es más alto que si se utilizan en cuestionarios de evaluación continua. Además, al ser un tema más empresarial, no se disponía de información en la red con ejemplos o detalles concretos.

5. BIGE (Biogeografía):

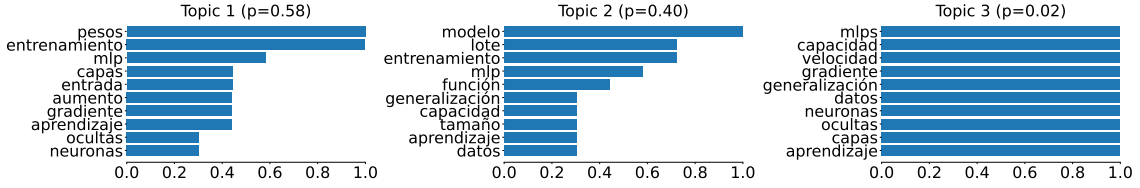
   - En esta asignatura se generaron preguntas con contexto de diferentes temas con conceptos básicos de clasificación de seres vivos, corología, fitogeografía y zoogeografía. Las preguntas se generaron con la intención de utilizarse como cuestionarios en clase, de control y también reservando algunas de ellas algunas de ellas para el examen final. La mayoría de las presentaciones en PDF que se utilizaron como contexto tenían más gráficos que texto, pero los títulos podían dar referencias al GPT-4 de la temática. Se observa que GPT-4 generó buenas preguntas que exploraban la distribución geográfica de especies y los factores que influyen en su dispersión, acorde con la temática; pero en algunos casos las preguntas resultaron ser variaciones de los mismos conceptos, excesivamente sencillas y repetitivas.

6. TRE (Teoría de Redes Eléctricas):

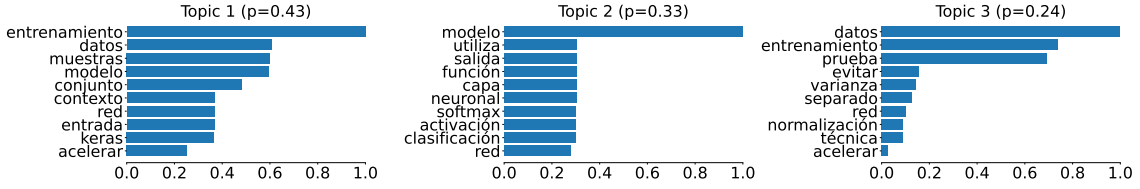
   - GPT-4 generó preguntas que abordaban conceptos avanzados de teoría de redes eléctricas, como el análisis de circuitos en corriente alterna con números complejos y los teoremas de redes. Se demostró la capacidad de GPT-4 para generar preguntas cuya respuesta requería la comprensión de los conceptos teóricos fundamentales para abordar de forma práctica la resolución de circuitos eléctricos.



(a) Preguntas planteadas por el profesorado, disponibles en Aula Virtual



(b) Preguntas generadas por GPT-4 (sin contexto)



(c) Preguntas generadas por GPT-4 (utilizando contexto)

Fig. 2: Análisis por temas latentes (topics) aplicando LDA sobre preguntas tipo test para la asignatura Modelos Conexionistas y el material de la sesión de laboratorio correspondiente al uso de redes neuronales utilizando la librería de aprendizaje profundo Keras (Python). Temas obtenidos para (a) las preguntas disponibles en Aula Virtual, planteadas por el profesorado (b) preguntas generadas por GPT-4 (sin contexto) y (c) preguntas generadas por GPT-4 (utilizando contexto, es decir, material en PDF disponible para la sesión de laboratorio). Los temas aparecen ordenados de mayor a menor importancia para cada uno de los conjuntos, señalando su probabilidad p.

4.2 Validación semántica del uso contextualizado de GPT-4

En segundo lugar, se realizó una validación semántica utilizando LDA para verificar el uso contextualizado de GPT-4. Para ello, se seleccionó como referencia una sesión de laboratorio de la asignatura MC, la cual se corresponde con el uso de la librería de aprendizaje profundo Keras (lenguaje de programación Python) para la implementación de redes neuronales. Para evaluar semánticamente el uso de GPT-4 para generar preguntas tipo test, se obtuvieron los tres temas latentes más representativos de los siguientes conjuntos: 1) las preguntas tipo test disponibles en Aula Virtual para esta práctica, planteadas por el profesorado; 2) las preguntas tipo test generadas por GPT-4 sin contexto; y 3) las preguntas tipo test generadas por GPT-4 con contexto, siendo el contexto el material PDF disponible para la realización de la práctica.

Los temas se muestran en la Figura 2, ordenados, para cada conjunto de preguntas, de mayor a menor importancia para su representación, e indicando su probabilidad p correspondiente. Si comparamos los temas que representan las preguntas planteadas por el profesorado disponibles en Aula Virtual con los temas para las preguntas generadas por GPT-4, utilizando o no contexto, se puede observar que, en general, GPT-4 es capaz de generar preguntas atendiendo a temas similares a los que consideraría el profesorado. Para los tres conjuntos de preguntas, los temas latentes aparecen representados por palabras como “entrenamiento”, “datos”, “capa” (“capas”,

"layers") o "modelo"(y sinónimos o palabras relacionadas como "parámetros", "pesos" o "red"). Esto demuestra que GPT-4 puede ser útil en general para la generación de preguntas tipo test. Además, si comparamos los temas que representan las preguntas generadas con o sin contexto, se observa que el uso contextualizado de GPT-4 permite obtener un conjunto de preguntas cuyos temas principales (y la proporción de los mismos) se acerca más a los planteados por el profesorado. Si bien no es posible la correspondencia perfecta entre los temas considerados por el profesorado y los asociados al uso contextualizado de GPT-4, sí que podemos observar probabilidades de ocurrencia similares para determinadas palabras. Por tanto, se puede concluir que el uso contextualizado de GPT-4 se acerca más a los resultados que esperaríamos del planteamiento de preguntas tipo test por el profesorado.

4.3 Análisis cuantitativo de las preguntas generadas

El análisis gráfico presentado en la Figura 3 muestra la distancia de Jaccard entre las preguntas generadas por GPT-4 y las versiones corregidas por la docente para la asignatura TRE. En dicha asignatura, compuesta por siete temas (del 0 al 6), se generaron diez preguntas por tema. Los resultados indican variaciones en la similitud de los textos entre diferentes temas.

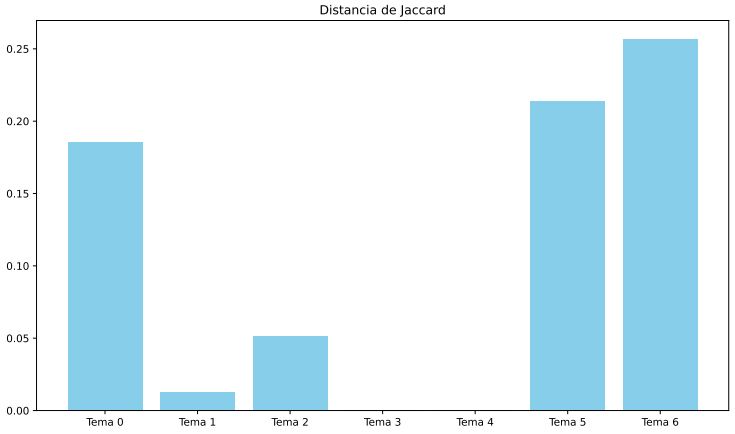


Fig. 3: Distancia de Jaccard entre el texto correspondiente a las preguntas generadas por GPT-4 y las correcciones por la docente para cada tema de la asignatura TRE.

Como se observa, los temas 0, 5 y 6 presentan las mayores distancias, lo que sugiere que las preguntas generadas por GPT-4 requirieron más correcciones para alinearse con la propuesta docente. Por el contrario, las diez preguntas generadas para cada uno de los temas 3 y 4 no precisaron de ningún tipo de corrección. Estos resultados pueden reflejar diferencias en la complejidad o en la especificidad del contenido de los temas.

Teniendo en cuenta todos los resultados de la asignatura, la media de la distancia de Jaccard fue de 0.10, lo que significa que alrededor del 90 % del contenido generado por GPT-4 fue de utilidad para generar los cuestionarios finales y no requirió ser corregido.

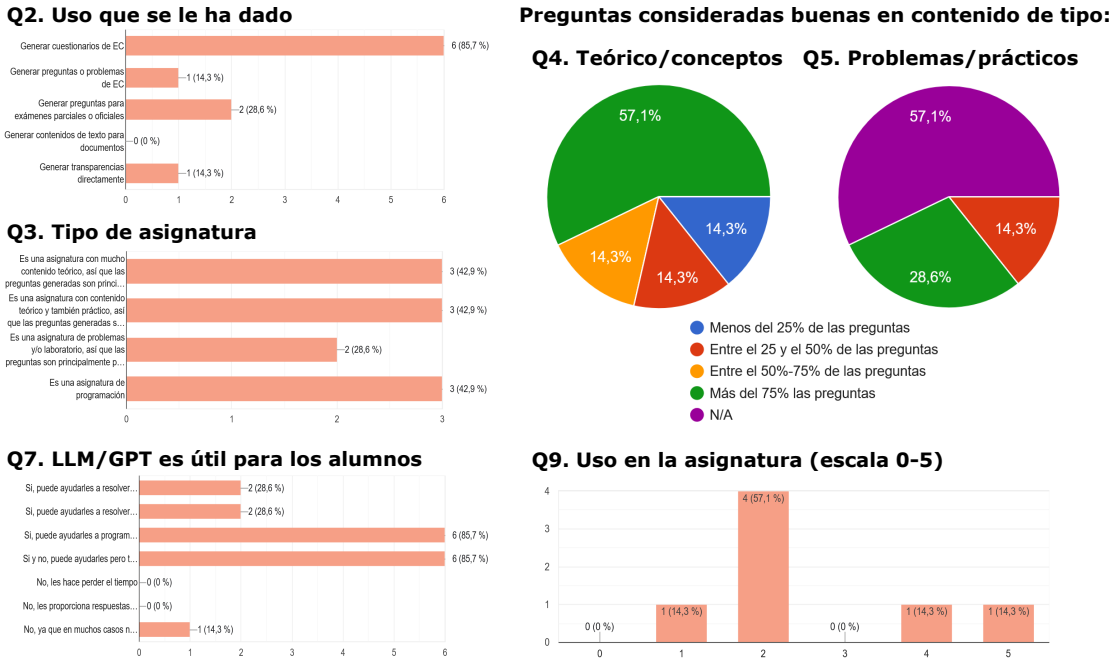


Fig. 4: Resultado de algunas preguntas de la encuesta al profesorado.

4.4 Encuestas al profesorado sobre el uso de LLMs en el aula

En la Figura 4 se muestran los resultados para las preguntas principales del Anexo A. En cuanto al uso que se le ha dado al modelo GPT-4 (cuestión Q2), destacar que mayoritariamente se ha usado para generar preguntas de evaluación continua y sólo en dos asignaturas, SIF y BIGE, se generaron preguntas para exámenes. Las asignaturas son normalmente mezcla de contenido teórico y práctico (Q3), destacando que tres de ellas requieren habilidades de programación. Si analizamos la cantidad de preguntas aprovechables en función de si el contenido es teórico o de conceptos (Q4), observamos que en 4 de 7 asignaturas más del 75 % son útiles. En cambio, sólo en 3 asignaturas se utilizaron para obtener cuestiones de problemas con éxito, correspondiéndose con ejercicios de programación (Q5). Cabe destacar también que todos los docentes han considerado el modelo GPT-4 útil para el docente (Q6), si bien en cuanto a generación de contenidos sólo dos asignaturas han hecho un uso significativo (Q9). En cuanto a la utilidad de GPT-4 para los alumnos (Q7), los docentes mayoritariamente consideran que puede ser una herramienta de apoyo en el proceso de aprendizaje, aunque todos afirman a su vez que puede generar respuestas incorrectas o parcialmente correctas. Por lo tanto, planteamos que el uso de modelos generativos como GPT-4 debe ir siempre acompañado de un razonamiento crítico.

Finalmente, la última pregunta de la encuesta (Q9) recoge las ventajas e inconvenientes de usar modelos generativos según la perspectiva del docente. Entre las conclusiones, se observó que GPT-4 permite generar contenido de manera rápida y, aunque generalmente es útil, a menudo necesita modificaciones, lo cual requiere un tiempo adicional. El uso de contexto mejoró en todo caso el resultado obtenido. Se observó también que la generación de un gran número de preguntas produce respuestas que tienden a ser similares o demasiado simples.

5 Conclusiones

Este estudio ha demostrado el potencial de GPT-4, un modelo generativo de inteligencia artificial, en la generación de preguntas de evaluación para diversas asignaturas que se imparten en los grados de Ciencia de datos, Ingeniería Electrónica Industrial y Geografía y Medio ambiente en la Universitat de València. A través de un enfoque metódico y una evaluación cuantitativa y cualitativa, hemos observado que GPT-4 facilita significativamente el proceso de creación de cuestionarios, proporcionando una herramienta valiosa tanto para docentes como para estudiantes. El modelo GPT-4 muestra una gran capacidad para generar preguntas contextualizadas y relevantes, especialmente cuando se alimenta con material específico de las asignaturas. Esta capacidad no solo optimiza el tiempo dedicado a la elaboración de cuestionarios de evaluación por parte del docente, sino que también enriquece la calidad del contenido. El uso de la inteligencia artificial generativa abre puertas a la personalización del aprendizaje, permitiendo la creación de cuestionarios adaptados a las necesidades y al progreso individual de cada estudiante. Sin embargo, es importante reconocer las limitaciones observadas en el estudio. Los resultados mostraron diferencias en la satisfacción de los docentes con las preguntas generadas entre las materias estudiadas. Se observó que para materias más básicas o de conocimiento genérico, así como aquellas más estrechamente vinculadas con las nuevas tecnologías, programación o ciencia de datos, los resultados fueron más positivos. Este aspecto pone de manifiesto que aquellas materias para las que existen más recursos en la red en abierto, son más proclives a obtener buenos resultados, aun cuando el modelo GPT-4 es alimentado con información específica de la asignatura a través de documentos, guiones de prácticas o diapositivas en formato PDF.

Este estudio subraya la necesidad de una supervisión final por parte del docente de los contenidos generados por GPT-4. A pesar de la eficiencia y la precisión de este modelo en la generación de cuestionarios, la intervención humana es y seguirá siendo crucial para garantizar la pertinencia y la corrección completa del material generado. Esto subraya la complementariedad entre la inteligencia artificial y la experiencia humana, en lugar de sugerir un reemplazo de una por la otra.

Referencias bibliográficas

Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Goldberg, Y. (2022). *Neural network methods for natural language processing*. Springer Nature.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull Soc Vaudoise Sci Nat*, *37*, 547-579.

Ribera, M., & Díaz Montesdeoca, O. (2024). ChatGPT y educación universitaria. Posibilidades y límites de ChatGPT como herramienta docente. *Octaedro*.

Skinner, B. F. (1958). Teaching Machines. *Science*, *128*(3330), 969-977.

Wilkinson, D. L. (2002). The Intersection of Learning Architecture and Instructional Design in e-Learning. En E. S. Series (Ed.), *e-Technologies in Engineering Education: Learning Outcomes Providing Future Possibilities* (pp. 213-221).

A Anexo: Encuesta sobre el uso de GPT-4 en el aula

Este anexo incluye el cuestionario que se facilitó al profesorado.

Preguntas:

Q1. Indica la asignatura donde se ha usado GPT-4: [Formato: Asignatura, curso, grado]:

Q2. Indica el uso que se le ha dado a la herramienta GPT-4 (selección múltiple):

- Generar cuestionarios de EC
- Generar preguntas o problemas de EC
- Generar preguntas para exámenes parciales o oficiales
- Generar contenidos de texto para documentos
- Generar transparencias directamente
- Otro: ...

Q3. Indica el tipo de asignatura (selección múltiple):

- Es una asignatura con mucho contenido teórico, así que las preguntas generadas son principalmente de conceptos.
- Es una asignatura con contenido teórico y también práctico, así que las preguntas generadas son mixtas (conceptos y problemas)
- Es una asignatura de problemas y/o laboratorio, así que las preguntas son principalmente problemas.
- Es una asignatura de programación

Q4. Al generar las preguntas de contenido teórico o conceptos, cuántas preguntas generadas suelen ser buenas para usarse directamente:

- Menos del 25 % de las preguntas
- Entre el 25 y el 50 % de las preguntas
- Entre el 50 %-75 % de las preguntas
- Más del 75 % las preguntas
- N/A (No se aplica)

Q5. Al generar las preguntas de contenido práctico o problemas, cuantas preguntas generadas suelen ser buenas para usarse directamente:

- Menos del 25 % de los casos
- Entre el 25 y el 50 % de los casos
- Entre el 50 %-75 % de los casos
- Más del 75 % de los casos
- N/A (No se aplica)

Q6. ¿Consideras que los modelos generativos como GPT-4 son una herramienta útil para el docente? Sí/No

Q7. ¿Consideras que los modelos generativos como GPT-4 son una herramienta útil para los alumnos (selección múltiple)?

- Sí, puede ayudarles a resolver dudas teóricas
- Sí, puede ayudarles a resolver ejercicios de problemas
- Sí, puede ayudarles a programar mejor
- Sí y no, puede ayudarles pero también dar respuestas incorrectas o parcialmente correctas.
- No, les hace perder el tiempo
- No, les proporciona respuestas erróneas en muchos casos
- No, ya que en muchos casos no saben distinguir las respuestas incorrectas o parcialmente correctas.

Q8. En una escala del 0 (nada) - 5 (mucho), ¿cuánto consideras que has usado los modelos generativos para generar contenidos o preguntas en la asignatura?

Q9. En tu asignatura, explica brevemente las ventajas e inconvenientes del uso que has realizado de los modelos generativos.