educade
Revista de Educación en Contabilidad, Finanzas y Administración de Empresas
The Spanish Journal of Accounting, Finance and Management Education

Nº 16 · 2025 ·
ISSN 2173-478X
pp. 83-95

Modelos de IA generativa para el soporte de preguntas frecuentes en cursos universitarios

Generative AI models for supporting frequently asked questions in university courses

Carla Antonini [1] (carla.antonini@uam.es) Corresponding author
https://orcid.org/0000-0001-7139-2249

Iván González [2] (ivan.gonzalez@uam.es)
https://orcid.org/0000-0002-5886-240X

Jorge E. López de Vergara [2] (jorge.lopez_vergara@uam.es)
https://orcid.org/0000-0002-4057-4688

Francisco J. Gómez-Arribas [2] (francisco.gomez@uam.es)
https://orcid.org/0000-0001-5363-171X

Aythami Morales [2] (aythami.morales@uam.es)
https://orcid.org/0000-0002-7268-4785

Rosa M. Carro [3] (rosa.carro@uam.es)
https://orcid.org/0000-0001-9684-5179

Álvaro Ortigosa [3] (alvaro.ortigosa@uam.es)
https://orcid.org/0000-0002-7674-4132

Eloy Anguiano [3] (eloy.angiano@uam.es)
https://orcid.org/0000-0003-4258-3374

José Luis Ucieda [1] (joseluis.ucieda@uam.es)
https://orcid.org/0000-0002-4473-5688

Luis de Pedro [2] (luis.depedro@uam.es)
https://orcid.org/0000-0002-4595-7370

(1) Departamento de Contabilidad, Universidad Autónoma de Madrid, (España).
(2) Departamento de Tecnología Electrónica y de las Comunicaciones, Universidad Autónoma de Madrid, (España).
(3) Departamento de Ingeniería Informática, Universidad Autónoma de Madrid, , (España).

https://dx.doi.org/10.12795/EDUCADE.2025.i16.06

Experiencia docente. Recibido: 23-09-25 – Revisado 20-10-25; Aceptado: 06-11-25

Resumen: Este trabajo presenta el resultado de un proyecto de innovación docente realizado en varios departamentos universitarios. El proyecto está orientado al diseño y desarrollo de una metodología propia de esta universidad que permita a cualquier docente generar un sistema apoyado en Inteligencia Artificial y validación humana para la consulta rápida de preguntas relativas a una asignatura. El beneficio del sistema sería ayudar al docente a responder preguntas simples típicas de una base de datos tipo Preguntas Frecuentes o *Frequently Asked Questions* (FAQ) para poder concentrar su tiempo de atención al estudiantado en tutorías de valor añadido.

Palabras clave: Inteligencia Artificial Generativa, Contexto, Exámenes de Opción Múltiple, Preguntas Frecuentes.

Abstract: This paper presents the results of a teaching innovation project carried out across several university departments. The project is aimed at designing and developing a methodology specific to this university that enables any faculty member to create a system supported by Artificial Intelligence and human validation for the quick consultation of questions related to specific courses. The benefit of the system would be to assist instructors in answering simple, frequently asked questions (FAQ)—similar to a FAQ database—so they can focus their time on providing value-added tutoring and personalized student support.

Keywords: Generative Artificial Intelligence, Context, Multiple-Choice Exams, Frequently Asked Questions.

1. INTRODUCCIÓN

La utilización de técnicas de la denominada frecuentemente como inteligencia artificial (IA, o *AI* por sus siglas en inglés) para la generación automática de textos es un tema de actualidad, tanto en el entorno académico como en la sociedad en general (p. ej. ChatGPT).

Aunque las tecnologías detrás de las implantaciones existentes de sistemas de este tipo son conocidas desde hace tiempo, es ahora cuando el rendimiento de los modernos procesadores paralelos *Graphics Processing Units* (GPU) hacen viable su aplicación en entornos interactivos, con tiempos de respuesta razonables, en torno a segundos.

Se han publicitado iniciativas para gestionar y controlar aspectos de la utilización de los generadores de IA de texto (ChatGPT y similares) que impactan directamente en la docencia. En relación con esta última herramienta, la literatura destaca que los alumnos utilizan ChatGPT sobre todo para el aprendizaje de idiomas, educación en línea, programación o codificación, redacción y traducción, aprendizaje personalizado, mientras que los docentes lo utilizan para el diseño de tareas, evaluaciones y exámenes, la supervisión de estudiantes, los planes de clase y la redacción de instrucciones (Baig & Yadegaridehkordi, 2024).

Siguiendo con el apoyo a la docencia, hay una amplia investigación explorando sistemas de tutoría inteligente. Hay evidencia de que estas herramientas tienen una gran variedad en cuanto a sus dominios de tarea, interfaces de usuario, estructuras de software, bases de conocimiento, etc. (Van Lehn, 2025). Sin embargo, hay cierto consenso en su efectividad para reemplazar ciertas tareas docentes tales como la asignación de deberes, preparación de preguntas para dinamizar la clase, respuestas a preguntas frecuentes u otras tareas análogas (Kaliisa et al., 2025; Gaskin et al., 2024; Van Lehn, 2011). Los modelos más populares para programar asistentes virtuales hasta la fecha son BERT, seguido por los modelos GPT-3, T5 y GPT-3.5 y sirven sobre todo para diseñar asistentes virtuales para tareas como son la generación de preguntas, la

calificación de respuestas y la corrección y explicación de código (García-Mendéz, 2025).

El proyecto de innovación descrito en este trabajo se orienta a la utilización de estas herramientas de tutorización programada con IA para facilitar la labor docente, aprovechando su aspecto fiable, limitando la entrada de datos a fuentes contrastadas: bibliografía de la asignatura y a preguntas contextualizadas.

Este trabajo, además, ha permitido que dicha universidad disponga de una metodología contrastada de apoyo en la docencia basada en la IA. Dicha metodología está alineada con los principios de innovación y excelencia en la docencia buscados por la Universidad, incorporando metodologías activas que faciliten la consecución de los resultados de aprendizaje en el estudiantado a través de herramientas TIC.

2. OBJETIVOS

El objetivo general del proyecto ha sido generar una aplicación tipo *chatbot* que permita al estudiantado aprender de una manera autónoma, personalizada, y adaptada a sus necesidades. Cada estudiante así podrá formular sus preguntas en el momento y de la manera que desee y obtener respuestas de alta calidad que se adaptan a su nivel de conocimientos, necesidades específicas de aprendizaje, e incluso recibir retroalimentación y consejos sobre cómo graduar el aprendizaje de cualquier tema o concepto de la asignatura de una manera inmediata.

Los objetivos específicos del proyecto han sido:

(1) Desarrollar un modelo transformacional para consultas sobre una asignatura.
(2) Definir una metodología para el desarrollo del modelo transformacional aplicable a diferentes asignaturas.
(3) Instalar el modelo en la infraestructura del grupo de investigación o, llegado el caso, de dicha universidad.
(4) Publicar el módulo en cursos de Moodle mediante un enlace para permitir al estudiantado acceder el mismo.
(5) Poner a disposición de la comunidad docente universitaria la metodología desarrollada.

La idea de utilizar la IA como tutor ya se ha sugerido anteriormente (Mollick, 2024), aunque de forma genérica. Este proyecto pretende ir más allá y entrenar un LLM (*Large Language Model*, Modelo Grande de Lenguaje) de manera específica en distintas asignaturas para proporcionar respuestas más precisas y una retroalimentación de más calidad al estudiantado.

3. DESCRIPCIÓN DE LA EXPERIENCIA

3.1. Recursos a disposición

Los recursos necesarios para llevar a cabo el proyecto han sido los siguientes:

(1) Conocimiento: el equipo del proyecto posee los conocimientos y la experiencia necesarios para su desarrollo. En el pasado inmediato se han abordado proyectos similares en la industria con éxito.
(2) Recursos: El grupo de investigación posee la infraestructura necesaria para poder realizar el desarrollo y las pruebas correspondientes sin necesidad de

financiación adicional. Adicionalmente, la Escuela Politécnica de dicha universidad ha puesto en marcha recientemente un clúster de GPUs para docencia.

(3) Datos: El modelo propuesto se ha basado en la utilización de preguntas de exámenes de opción múltiple contextualizados en la bibliografía de la asignatura. El equipo docente dispone de abundantes exámenes (en algún caso, acumulados durante casi treinta años de docencia) que permiten su utilización en el proyecto tras su correspondiente adaptación al modelo.

3.2. Método y desarrollo

Esta subsección se dividirá en dos partes principales. En la primera, se presentará información detallada sobre el proyecto de innovación docente, incluyendo sus objetivos, centros y participantes. En la segunda parte, se abordará el desarrollo del modelo de tutor virtual, describiendo su diseño, implementación y las herramientas tecnológicas empleadas para su construcción.

Método de implantación del proyecto de innovación docente

El objetivo general del proyecto es desarrollar y probar un sistema de generación de módulos de consulta de primer nivel para el estudiantado de una asignatura determinada. Este sistema estaría basado en modelos transformacionales o de atención que sería propiedad de la universidad y podrían ser utilizados en diferentes disciplinas, sin necesidad de pagar licencias de software a entidades externas. El proyecto ha tenido una duración de un curso académico, y han participado dos centros de la universidad: la Escuela Politécnica Superior y la Facultad de Ciencias Económicas y Empresariales, participando un total de 10 docentes de tres departamentos distintos, cubriendo un total de 6 asignaturas distintas. De éstas, hemos seleccionado algunos casos para compararlas en este artículo.

Desarrollo

En estos últimos dos años, los modelos grandes de lenguaje o LLMs, como GPT ("*Generative Pre-trained Transformer*" (Chemmalar Selvi, 2024)) y BERT (*Bidirectional Encoder Representations from Transformers*) (Chang et al., 2019), han transformado la forma en que interactuamos con la tecnología y están abriendo nuevas posibilidades en campos como la traducción automática, el análisis de textos, y en el caso particular de la docencia universitaria, ofrecen un sinfín de posibilidades relacionadas con las actividades docentes tanto a docentes como a estudiantes. Estos modelos están basados en lo que se conoce como inteligencia artificial generativa, IA generativa (Feuerriegel et al., 2024) y pueden crear contenido nuevo y plausible a partir de una serie de datos de entrada. Para ello, se entrenan con enormes conjuntos de datos, lo que permite al modelo aprender a predecir la siguiente palabra en una secuencia de texto basándose en las palabras anteriores. Este proceso se repite millones de veces hasta que el modelo desarrolla una comprensión profunda del lenguaje.

Si bien el uso de estos modelos mediante APIs o la web se ha generalizado bastante, la posibilidad de crear un modelo con nuestros propios datos presenta varios retos, como la necesidad de disponer de un conjunto amplio de datos propios para entrenar el modelo, disponer de recursos hardware de altas prestaciones para dicho entrenamiento, y posteriormente, controlar que el modelo no dé resultados incorrectos o "alucine". Como alternativa al entrenamiento de los modelos, y también para reducir las alucinaciones de los LLMs, ha surgido recientemente una alternativa conocida como Generación Aumentada con Recuperación (RAG, *Retrieval-Augmented Generation*)

(Sefika & Adrian, 2024) que consiste en ampliar el conocimiento de un modelo ya existente con un conjunto de datos que actúe como contexto de la respuesta del modelo, de modo que sea posible "dirigir" las respuestas del modelo. El sistema RAG funciona como un examen de libro abierto, integrando información relevante, obtenida de una base de datos vectorial, directamente en la consulta.

La clave para el correcto funcionamiento del RAG reside en la elección del contenido, dado que es necesario que exista similitud de los contenidos con las preguntas a realizar. Como se indicó, estos contenidos se almacenan en una base de datos vectorial, un tipo de base de datos diseñada para manejar incrustaciones de vectores, representaciones de datos de alta dimensión, usadas en el aprendizaje automático. Estas incrustaciones pueden representar varios tipos de datos, incluidos texto, imágenes y audio (Bao et al., 2024). Las bases de datos de vectores están optimizadas para la búsqueda de similitudes, lo que permite la recuperación rápida de los vectores más parecidos basándose en métricas de distancia como la similitud del coseno o la distancia euclídea.

Por ello, en este proyecto seguimos la metodología RAG, como se muestra en la figura 1, y usaremos un modelo LLM potente y ya entrenado como Mixtral-8x7B (Mistral, 2025; Huggingface, 2025), al que complementaremos con una base de datos vectorial Chroma (Chroma, 2025) con contenido de las diferentes materias a las que queremos que el modelo de soporte. De ese modo, podemos guiar al modelo para que responda como un experto en dichas materias. Estos textos pueden ser partes de un libro, de una presentación, un conjunto de preguntas y respuestas, etc.

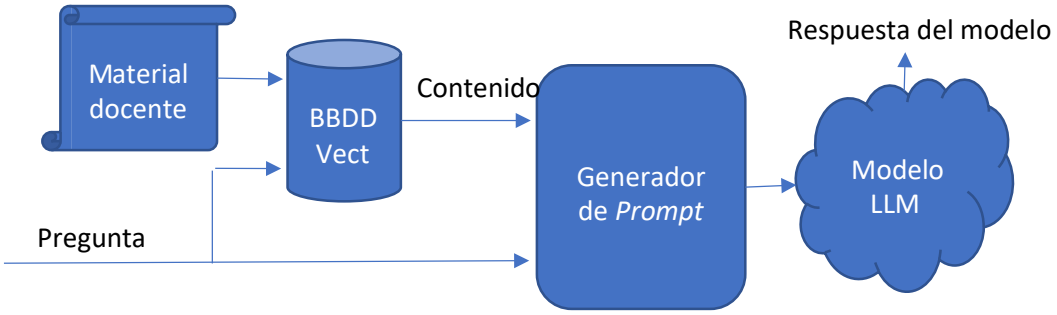


Figura 1. Arquitectura RAG.

Para probar la metodología hemos seleccionado las materias en las que los investigadores de este proyecto imparten docencia: redes de comunicaciones, arquitectura de computadores, sistemas operativos y análisis de estados financieros. A cada docente se le ha pedido un conjunto de preguntas de su materia, y la documentación que el estudiantado necesitaría para poder responder dichas preguntas (fragmentos de libros, presentaciones, etc.). Como es habitual en este tipo de situaciones, las preguntas se encuentran en diferentes formatos: Word, LaTeX, MoodleXML, PDF, etc. por lo que se han desarrollado diferentes herramientas para obtener los contenidos de estos formatos, y que sea posible la carga de estos contenidos en la base de datos vectorial. En el caso de las preguntas y respuestas, también es posible emplearlas para evaluar el modelo, como se verá más adelante. Una vez se dispone de los contenidos cargados en la base de datos, el siguiente paso ha sido el desarrollo de la aplicación tipo *chatbot* que usará el RAG (modelo + base de datos) para responder las preguntas. Ya que en una primera fase queremos evaluar la eficiencia del modelo, se ha desarrollado también una versión para validación del RAG

que permite plantear las preguntas al RAG de forma automática para evaluar los resultados, en vez de emplear el *chatbot*.

4. RESULTADOS

Para evaluar el RAG propuesto, se ha cargado la base de datos solo con los contenidos ofrecidos al estudiantado y se ha enviado al RAG la batería de preguntas de cada materia. En una primera versión del *prompt* (instrucciones que se proporcionan a la IA) se usa únicamente la pregunta, que puede incluir las posibles respuestas en caso de ser multiopción, y el conjunto de textos almacenados en la base de datos más similares a la pregunta, tal y como se muestra en la figura 1. En este caso, el RAG responde a la pregunta empleando su "propio conocimiento" y los contenidos obtenidos como contexto. Este escenario se corresponde con el propuesto como objetivo del proyecto, que es responder preguntas del estudiantado a través del *chatbot*. Los resultados de estas pruebas muestran que el modelo suele responder correctamente a las respuestas multi-opción, y tiene más dificultades para responder a preguntas de rellenar, como se verá en los apartados siguientes. En el caso de las respuestas multi-opción, incluso justifica cada una de las posibles respuestas. Lo que hemos podido comprobar es que a veces se equivoca en la elección de la respuesta, pero justifica la elección añadiendo información que hace que la respuesta sea correcta. Esto se debe a que habitualmente se ponen respuestas incompletas, y el modelo elige la respuesta correcta completando lo que falta. Esto nos lleva a pensar que se puede utilizar el RAG no solo para responder a preguntas, sino para validar la calidad de las preguntas de un examen, por ejemplo.

En una segunda versión del *prompt*, se ha ampliado el *prompt* anterior incluyendo la respuesta correcta, de modo que se pide al modelo que justifique si ésta le parece o no correcta. En estos casos suele apostar por la respuesta correcta, y da el razonamiento para aceptarla. También suele justificar por qué las otras opciones no son correctas, en caso de preguntas multi-opción. Este escenario nos permite obtener una respuesta ampliada que podemos usar posteriormente para ampliar la base de datos vectorial, en vez de emplear simplemente la pregunta y respuesta. Obviamente, se hace necesario que el personal docente valide la explicación para evitar errores en las respuestas al estudiantado.

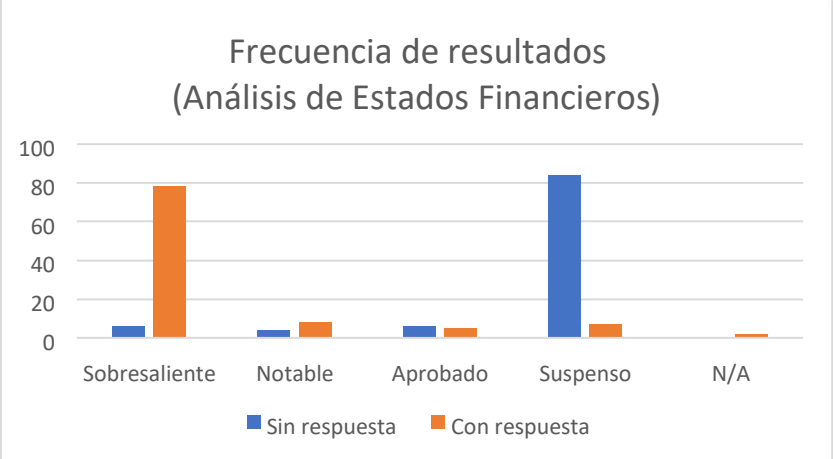
4.1 Análisis de Estados Financieros

En la asignatura de 'Análisis de Estados Financieros' se han utilizado preguntas de respuesta múltiple con tres alternativas posibles. Las cuestiones se han obtenido del banco de preguntas de Moodle que se viene utilizando en la asignatura por su equipo docente desde hace varios años. En general, se utiliza para generar cuestionarios de 15-20 preguntas generadas aleatoriamente de cada tema para que los estudiantes puedan evaluar su aprendizaje de una manera rápida y sencilla. Las preguntas incluyen comentarios de retroalimentación para las respuestas incorrectas que explican cuál es la correcta y por qué.

Un análisis de 100 preguntas tomadas aleatoriamente del banco de preguntas permite observar que el modelo responde mejor cuando tiene la respuesta correcta, esto es, cuando posee más contexto (ver Figura 2). Cuando no posee las respuestas correctas, 84 respuestas fueron incorrectas y de los 16 restantes, solo 6 pueden considerarse completamente correctas. En general, se observa que el modelo responde mejor a preguntas generales, de conocimientos básicos, de definir y recordar conceptos, pero tiene dificultades con las preguntas de relacionar, aplicar, o calcular. Para ello, necesitaría *prompts* más elaborados. Por ejemplo, cuando se pregunta qué opción (de las tres que tiene cada pregunta) permite medir la ratio ROE, el modelo responde

definiendo la ROE sin indicar cuál es la respuesta correcta. En otras, cuando se le pregunta por el efecto en los estados financieros de la contabilización de una cuota de un préstamo por el método francés, el modelo responde que cada cuota incluye una parte de intereses y otra de amortización del capital. Esta respuesta es correcta, pero no responde a la pregunta.

Figura 2. Resultados obtenidos con preguntas de Análisis de Estados Financieros



Cuando se le proporcionan las respuestas correctas al modelo, las proporciones se invierten. De las 100 preguntas analizadas, 78 fueron correctas y solo 7 se etiquetaron incorrectas. De nuevo observamos el mismo problema que en el caso de sin respuestas. Por ejemplo, cuando se le pregunta cómo se contabilizaría un anticipo de un cliente, el modelo opta por registrarlo –incorrectamente– como un activo corriente, justificando además que el epígrafe 'Clientes y Anticipos de Clientes' es un recurso disponible. En otra pregunta sobre si la ratio MTB (*Market-To-Book*) puede ser negativa, el modelo asume que MTB es — incorrectamente— la 'Media Ponderada de Capital', posiblemente al traducir incorrectamente el término.

Un análisis más detallado permite observar los sesgos del modelo, es decir, de los datos con los que ha sido entrenado. Por ejemplo, varias respuestas eran en inglés o parcialmente en inglés. En otros casos, la respuesta era incorrecta según la normativa contable española, pero correcta para la normativa contable de países anglosajones (como Estados Unidos, Reino Unido, etcétera). También se observan a veces expresiones y términos relacionados con la contabilidad y finanzas utilizadas en Latinoamérica (ganancias en vez de beneficios, por ejemplo).

4.2 Arquitectura de ordenadores

En la asignatura 'Arquitectura de Ordenadores' se han utilizado preguntas abiertas con respuesta corta, donde se valora la respuesta redactada con claridad, precisión y concreción, con menos de 150 palabras. Las preguntas se han sacado de la colección histórica de los exámenes de la asignatura. El contexto de información suministrada al modelo ha sido el contenido de 4 libros de texto en inglés, pero las preguntas se van a formular en castellano. La metodología de trabajo ha sido preparar una plantilla de corrección similar a la que usan docentes de la asignatura y que sirve de rúbrica para la evaluación homogénea de las respuestas del estudiantado cuando la corrección se realiza por varios docentes. En una primera prueba se formularán las preguntas sin dar como contexto la plantilla de corrección, para que el modelo extraiga el conocimiento, principalmente de los libros de texto, y se compararán estos resultados con los obtenidos

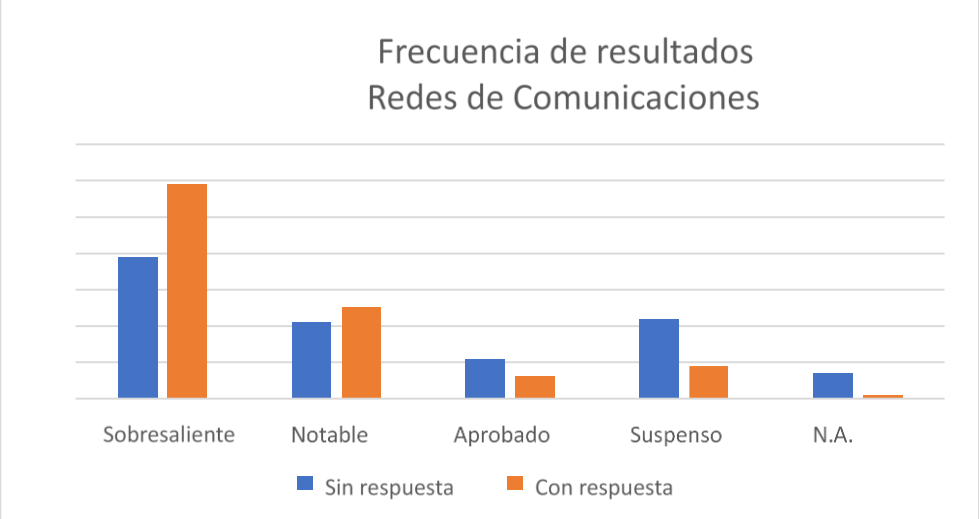
cuando se introduzca como contexto adicional la respuesta correcta. Los resultados se encuentran en el proceso de valoración por docentes expertos.

4.3 Redes de comunicaciones

En la asignatura de 'Redes de Comunicaciones' se han utilizado 100 preguntas de opción múltiple. Las preguntas se han tomado de la colección histórica de los exámenes de la asignatura. El contexto de información suministrada al modelo ha sido el contenido de 2 libros de texto en inglés, pero las preguntas se formulan en castellano. El profesor experto en la materia ha valorado las respuestas obtenidas, siguiendo un enfoque similar a (Hernández et al., 2025).

De manera aleatoria, se han omitido las posibles respuestas en unas preguntas, para ver cómo se comportaba el modelo generativo en ese caso. Como cabría esperar, en general, el modelo funciona mejor cuanto más contexto tenga para responder, por lo que el contar con las posibles respuestas ayuda a la generación de la respuesta. Cuando no se le proporciona respuesta, en algunos casos (8%) incluso ha ocurrido que el modelo la ha generado en inglés, pese a que ha sido configurado para que responda en castellano. Adicionalmente, el número de preguntas puntuadas con 0 (pregunta no válida) es mucho mayor cuando no se le proporciona la respuesta, al no saber el modelo cómo afrontar el enunciado de la pregunta.

Figura 3. Resultados obtenidos con preguntas de Redes de Comunicaciones



Según se muestra en la figura 3, como ya ocurría en casos previos, para dirigir el contenido de la respuesta resulta más útil que el modelo conozca cuál es la opción correcta (independientemente de que conozca las cuatro opciones). El modelo en este caso suele funcionar mejor, con menos preguntas no resolubles (N/A en la figura) y un mayor porcentaje de respuestas completamente satisfactorias. Las respuestas incorrectas suelen producirse cuando las preguntas se refieren a la resolución de una cuestión de carácter numérico. En estos casos (5 de las 9 incorrectas), aunque trata de explicar el método para resolver el problema, el modelo no es capaz de operar correctamente los resultados, inventándose los valores intermedios y finales para alcanzar la solución. Hay que destacar que las otras 4 respuestas incorrectas no son de

este tipo, por lo que su puntuación puede deberse a la aparición de alucinaciones en el modelo.

Como conclusión, podría decirse que el modelo, siempre que se conozca la respuesta, se puede comportar como una persona estudiante promedio, lo que permitiría validar preguntas de un examen antes de ponérselo al estudiantado. Por el contrario, se desaconseja su uso para resolución de dudas si no existe una supervisión previa por parte del personal docente antes de proporcionar la respuesta.

4.4. Análisis estadístico

Para evaluar la significación estadística de las diferencias observadas entre condiciones (con y sin proporcionar la respuesta correcta como contexto), se aplicaron dos enfoques estadísticos complementarios. Primero, se utilizó el test chi-cuadrado (χ^2) de Pearson para comparar las proporciones de respuestas correctas versus incorrectas entre condiciones, excluyendo las respuestas no evaluables (NA) del análisis. Se calcularon también el tamaño del efecto mediante la V de Cramer (McHugh, 2013) y los Odds Ratios (OR) (Peng, et al., 2010) para cuantificar la magnitud de las diferencias. Segundo, como análisis complementario, se aplicó el test no paramétrico de Mann-Whitney U considerando las puntuaciones ordinales (1 = correcta, 0.5 = parcial, 0 = incorrecta), lo que permitió aprovechar toda la información gradual de las respuestas. Los intervalos de confianza del 95% se calcularon para todas las proporciones estimadas. Todos los análisis se realizaron con a = 0.05. La Tabla 1 muestra los resultados de dichos tests.

Tabla 1. Análisis estadístico del rendimiento del sistema RAG por asignatura

Asign.	n	Sin contexto (% correct)	Con contexto (% correct)	χ^2	p	OR	V Cramer
[1]	100	6.0% [1.3-10.7]	79.6% [71.6-87.6]	106.75	<.001***	61.10	0.734
[2]	100	41.9% [31.9-52.0]	59.6% [49.9-69.3]	5.30	.021*	2.04	0.166

Notas: Los intervalos de confianza del 95% se muestran entre corchetes. OR = Odds Ratio; V Cramer = tamaño del efecto (V de Cramer). Todos los tests con gl = 1. El análisis excluye respuestas no evaluables (NA). *** p < 0.001; * p < 0.05. Asignaturas: [1] = Análisis de Estados Financieros; [2] Redes de Comunicaciones

Los análisis estadísticos confirmaron diferencias altamente significativas entre condiciones en ambas asignaturas evaluadas, aunque con magnitudes de efecto notablemente diferentes. En Análisis de Estados Financieros, la proporción de respuestas correctas aumentó significativamente de 6.0% [IC 95%: 1.3%-10.7%] sin contexto a 79.6% [IC 95%: 71.6%-87.6%] con contexto (χ^2 = 106.75, gl = 1, p < 0.001). El tamaño del efecto fue grande (V de Cramer = 0.734), y el modelo mostró 61 veces más probabilidad de responder correctamente cuando se proporcionó contexto adicional (OR = 61.10). El análisis complementario de Mann-Whitney confirmó estos hallazgos considerando las puntuaciones parciales (U = 9017.0, p < 0.001, r = 0.726), con una diferencia de medias de 0.75 puntos en la escala de 0-1. En Redes de Comunicaciones, aunque el efecto fue significativo, su magnitud fue considerablemente menor. La proporción de respuestas correctas aumentó de 41.9% [IC 95%: 31.9%-52.0%] a 59.6% [IC 95%: 49.9%-69.3%] (χ^2 = 5.30, gl = 1, p = 0.021), con un tamaño del efecto mediano (V de Cramer = 0.166) y un Odds Ratio de 2.04. El test de Mann-Whitney también confirmó diferencias significativas (U = 5613.5, p = 0.004, r = 0.189), con una diferencia de medias de 0.16 puntos. El análisis chi-cuadrado con tres categorías (correcta/parcial/incorrecta) reveló que el contexto no solo aumenta las respuestas correctas, sino que reorganiza completamente la distribución de respuestas en ambas asignaturas (Análisis Financiero: χ^2 = 127.25, p < 0.001; Redes: χ^2 = 9.37, p < 0.01). La notable diferencia en magnitud del efecto entre

asignaturas sugiere que el sistema RAG es especialmente efectivo en dominios alejados del conocimiento base del modelo (como la contabilidad española), donde el *baseline* sin contexto es muy bajo (6%). En dominios que el modelo ya conoce razonablemente (como redes de comunicaciones), el contexto adicional sigue siendo beneficioso, pero el efecto relativo es menor, posiblemente porque el modelo puede apoyarse en su conocimiento previo. Esta interpretación es consistente con la función del RAG como mecanismo de aumentación del conocimiento del modelo.

4.5. Resultados adicionales

Con el objetivo de evaluar el rendimiento de nuestro sistema frente a soluciones comerciales, se han llevado a cabo cuatro pruebas adicionales utilizando una de las asignaturas de referencia con ChatGPT versión 4o. Las pruebas se han diseñado de la siguiente forma:

(1) Prueba 1: Se suministró al modelo ChatGPT-4o el listado de 100 preguntas sin respuesta, solicitándole que generara las respuestas correspondientes.
(2) Prueba 2: Se proporcionó el listado de 100 preguntas con respuesta, pidiéndole al modelo que explicara las respuestas dadas.
(3) Prueba 3: Se repitió la prueba 1, pero usando una instancia personalizada de ChatGPT que había sido entrenada con la documentación de la asignatura (en una estrategia similar al uso de RAG en nuestra implementación).
(4) Prueba 4: Se replicó la prueba 2 con el modelo personalizado anteriormente mencionado.

Los resultados obtenidos fueron desiguales, pero en general inferiores al sistema propio que hemos desarrollado. A continuación, se destacan los principales hallazgos:

Primero, hemos recabado evidencia de ciertas limitaciones de ChatGPT en el manejo de listados extensos. En particular, se detectó un patrón recurrente en el que el modelo respondía correctamente a las primeras cinco preguntas, pero comenzaba después a inventar preguntas no presentes en el listado original. Segundo, se pudo observar una desincronización entre preguntas y respuestas: En el caso 4 en que se pidió al modelo explicar respuestas existentes, se observó que con frecuencia la numeración de las preguntas y sus explicaciones no coincidía sobre el archivo intercambiado, lo que afecta significativamente a la usabilidad del sistema. Tercero, si bien se observan mejoras con personalización, no son suficientes. El uso de un ChatGPT personalizado con la documentación de la asignatura (pruebas 3 y 4) produjo respuestas de mayor calidad, más alineadas con el contenido académico. Sin embargo, persistieron los problemas de invención y desalineación, especialmente en interacciones con listados largos.

4. DISCUSIÓN Y CONCLUSIONES

El uso de modelos LLM se ha extendido ampliamente entre la comunidad universitaria (p. ej. *ChatGPT*). En este escenario debe ser el personal docente o el estudiantado quien valide la exactitud del contenido generado por estos modelos de IA generativa.

Para asegurar que los resultados obtenidos de un modelo LLM son correctos, se puede optar por entrenar el modelo con contenidos propios, pero para ello es necesario disponer de datos y recursos computacionales. La alternativa es el RAG, que consiste en usar un modelo y complementar su conocimiento con un conjunto de datos que permitan guiar sus respuestas. De esta manera nos ahorramos el tiempo y coste del entrenamiento, y resulta más sencillo de ampliar las posibilidades de los modelos. En este

proyecto se desarrolló un RAG para validar el potencial de modelos LLM que permitan responder a las preguntas del estudiantado usando contenidos previamente seleccionados por los docentes. Así, se garantiza que las respuestas son razonablemente correctas, y evitamos que el modelo dé respuestas desactualizadas o incorrectas. Además, el uso del RAG también permite validar que las preguntas de un examen y sus respuestas pueden ser respondidas con la información aportada al estudiantado.

La evaluación del sistema RAG reveló mejoras significativas con contexto en las dos asignaturas universitarias analizadas en el proyecto, aunque con magnitudes contrastables. En Análisis de Estados Financieros, el rendimiento aumentó dramáticamente de 6.0% a 79.6% de acierto (+73.6 puntos porcentuales, $\chi^2 = 106.75$, $p < 0.001$, OR = 61.10), mientras que en Redes de Comunicaciones la mejora fue más modesta, de 41.9% a 59.6% (+17.7 puntos porcentuales, $\chi^2 = 5.30$, $p = 0.021$, OR = 2.04). Los tamaños del efecto fueron extraordinariamente grande (V de Cramer = 0.734) y mediano (V = 0.166) respectivamente, confirmados mediante análisis complementarios con el test de Mann-Whitney, lo que refuerza la robustez de los hallazgos.

Esta diferencia en magnitud del efecto no constituye una limitación del sistema, sino que identifica su nicho de máximo valor: dominios especializados fuera del conocimiento base del LLM. La contabilidad española, con normativa específica y escasa representación en corpus generalistas, requiere críticamente del RAG (modelo sin contexto esencialmente no funcional), mientras que en redes de comunicaciones el contexto mejora un rendimiento ya competente. Los resultados validan que la supervisión docente es imprescindible —con tasas de error del 20% incluso con contexto en Análisis Financiero— y confirman que estos sistemas aportan mayor valor diferencial en asignaturas de contenido especializado, normativa local o áreas poco representadas en corpus de entrenamiento, ofreciendo una solución escalable para instituciones que requieren adaptación a contextos específicos. De manera adicional, se comparó el sistema propio con instancias de ChatGPT-4o: los modelos comerciales tendieron a inventar ítems y a desincronizar preguntas y explicaciones en listados largos; la personalización mejoró la calidad, pero no eliminó esos fallos, mientras que la solución desarrollada (con contexto controlado y validación docente) resultó más fiable y coherente.

Por otro lado, los resultados del proyecto indican que el sistema basado en IA generativa funciona mejor en preguntas de opción múltiple y reconocimiento de conceptos cuando se le proporciona contexto adicional, como las respuestas correctas. Sin embargo, tiene dificultades con preguntas que requieren análisis, aplicación o cálculos. Por ello, las respuestas generadas por la IA deben ser supervisadas por un docente para evitar errores en las mismas. No obstante, pueden ser útiles para prever lo que una persona estudiante promedio respondería en un examen y evaluar así su dificultad, lo que añade valor al proceso educativo. Las pruebas adicionales con soluciones comerciales refuerzan la solidez de nuestra solución propia, especialmente en términos de fiabilidad y control de calidad en la generación de respuestas. Mientras que herramientas comerciales como ChatGPT ofrecen capacidades generales potentes, su rendimiento en contextos educativos específicos y estructurados presenta limitaciones que justifican el desarrollo de soluciones adaptadas.

Este trabajo presenta varias limitaciones que deben considerarse al interpretar los resultados. En primer lugar, la evaluación se realizó en solo dos asignaturas y una universidad, por lo que conviene replicar el método en más disciplinas y contextos para confirmar su generalidad. En segundo lugar, la validación carece de pruebas con estudiantes reales que midan impacto en aprendizaje, satisfacción y uso. Esta es sin duda una línea con grandes posibilidades para el futuro desarrollo de este trabajo. Por

último, el estudio empleó Mixtral-8x7B, que en el momento de llevar a cabo el estudio era una buena opción como LLM. Ensayar con LLMs más recientes podría mejorar los resultados. Aun así, los resultados muestran de forma robusta el valor del enfoque RAG con supervisión docente y orientan futuras líneas prácticas.

Los resultados abren múltiples líneas de investigación futura. En primer lugar, el chatbot puede evaluarse con uso real con estudiantes utilizando un diseño experimental, y medir el impacto en el aprendizaje, satisfacción y patrones de uso, por ejemplo. La rápida evolución de la tecnología invita a comparar la metodología con modelos LLMs más recientes y técnicas RAG avanzadas (*reranking*, *multi-hop*, detección de incertidumbre) y cuantificar mejoras. Una posible tercera línea sería ampliar el alcance pedagógico (tutoría socrática, diagnóstico adaptativo) y la validación en más disciplinas, incorporando métricas educativas y consideraciones de ética, privacidad y gobernanza institucional.

AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado por los siguientes proyectos de innovación docente de la UAM: FAQ-IA (EPS_008.23_INN), GPT4NETS (EPS_023.24_INN)

REFERENCIAS

Baig, M. I., & Yadegaridehkordi, E. (2024). ChatGPT in higher education: A systematic literature review and research challenges. *International Journal of Educational Research, 127*, 102411.

Bao, Z., Liao-Liao, L., Wu, Z., Zhou, Y., Fan, D., Aibin, M., Coady, Y., & Brownsword, A. (2024). *Delta Tensor: Efficient vector and tensor storage in Delta Lake*.

Chemmalar Selvi, G. (2024). *GPT (Generative Pre-Trained Transformer): A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. 12*, 54608–54649.

Chroma. (n.d.). *The AI-native open-source embedding database*. Recuperado el 16 de junio de 2024 de https://www.trychroma.com/

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT*, 4171–4186.

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). *Generative AI. Business & Information Systems Engineering, 66*(1), 111–126.

García-Méndez, S., de Arriba-Pérez, F., & Somoza-López, M. D. C. (2025). *A review on the use of large language models as virtual tutors. Science & Education, 34*(2), 877–892.

Gaskin, James; Blondeel, Eva; Schuetzler, Ryan; Serre, Rachel; Steffen, Jacob; and Wood, David (2024) *Chatbots Mitigate Help-Seeking Avoidance*. ICIS 2024 Proceedings. 14. https://aisel.aisnet.org/icis2024/learnandiscurricula/learnandiscurricula/14

Hernández, J. A., Conde, J., Reviriego, P., & Martínez Ruiz de Arcaute, G. (2023, July 25). *Is ChatGPT capable of solving classical communications and networking problems? TechRxiv*. https://doi.org/10.36227/techrxiv.23727174.v1

Hugging Face. (n.d.). *Mixtral-8x7B-Instruct-v0.1*. Recuperado el 16 de junio de 2024 de https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

Kaliisa, R., Misiejuk, K., López-Pernas, S., & Saqr, M. (2025). How does artificial intelligence compare to human feedback? A meta-analysis of performance, feedback perception, and learning dispositions. *Educational Psychology*, 1–32. https://doi.org/10.1080/01443410.2025.2553639

McHugh, M. L. (2013). The Chi-square test of independence. *Biochemia Medica*, 23(2), 143-149. https://doi.org/10.11613/BM.2013.018

Mistral AI. (n.d.). *Mixtral of experts.* Recuperado el 16 de junio de 2024 de https://mistral.ai/news/mixtral-of-experts/

Mollick, E. (2024). *Co-Intelligence: Living and working with AI.* Portfolio Penguin.

Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An Introduction to Logistic Regression Analysis and Reporting. *The Journal of Educational Research,* 96(1), 3–14. https://doi.org/10.1080/00220670209598786

Sefika, E., & Adrian, P. (2024). *Retrieval-augmented generation-based relation extraction.*

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education, 16*(3), 227–265.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197–221.