FAQ-IA: Soporte de Preguntas Frecuentes mediante la Inteligencia Artificial Generativa

Iván González[a], Jorge E. López de Vergara[a] , Francisco J. Gómez-Arribas[a] , Aythami Morales[a], Rosa Carro[b] , Álvaro Ortigosa[b] , Eloy Anguiano[b] , Carla Antonini[c] , José Luis Ucieda[c] , Luis de Pedro[a]

[a]Universidad Autónoma de Madrid, {ivan.gonzalez, jorge.lopez_vergara, francisco.gomez, aythami.morales, luis.depedro}@uam.es, Departamento de Tecnología Electrónica y de las comunicaciones.

[b]Universidad Autónoma de Madrid, {rosa.carro, alvaro.ortigosa, eloy.angiano}@uam.es, Departamento de Ingeniería Informática.

[c]Universidad Autónoma de Madrid, {carla.antonini, joseluis.ucieda}@uam.es, Departamento de Contabilidad.

Resumen

Esta contribución presenta el resultado de un proyecto de innovación docente (EPS_008.23_INN) realizado en la Universidad Autónoma de Madrid (UAM). El proyecto está orientado al diseño y desarrollo de una metodología propia de la UAM que permita a cualquier profesor o profesora de la Universidad 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.

FAQ-IA: FAQ Support through Generative Artificial Intelligence

This contribution presents the result of a teaching innovation project (EPS_008.23_INN) carried out at the Universidad Autonoma de Madrid (UAM). The project is oriented to the design and development of a methodology of the UAM that allows any university professor the generation of a system supported by Artificial Intelligence and human validation for the quick consultation of questions related to a subject. The benefit of the system would be to help the teacher to answer simple questions typical of a database of Frequently Asked Questions (FAQ) to concentrate his or her time with students on value-added tutorials.

Key words: Generative Artificial Intelligence, Context, Multiple Choice Tests, 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.e. 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. Existe preocupación en la comunidad académica sobre la facilidad que estas herramientas proporcionan al estudiantado para producir textos de aceptable calidad que puedan ser empleados para simular entregas solicitadas como evaluación de una asignatura.

El proyecto de innovación descrito en este trabajo se orienta en cambio a la utilización de estas herramientas para facilitar la labor docente, aprovechando su aspecto más fiable, limitando la entrada de datos a fuentes contrastadas (biografía de una asignatura) y a preguntas contextualizadas.

Este trabajo, además, ha permitido que la UAM 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 a los estudiantes 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 la UAM.
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 [1], 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 a los estudiantes.

3. 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 el desarrollo del mismo. En el pasado inmediato se han abordado proyectos similares en la industria con éxito.
2. Recursos: El grupo de investigación HPCN posee la infraestructura necesaria para poder realizar desarrollo y las pruebas correspondientes sin necesidad de financiación adicional. Adicionalmente, la EPS-UAM 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 equpo de profesores 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.

4. Desarrollo

En estos últimos dos años, los modelos grandes de lenguaje o LLMs, como GPT (*Generative Pre-trained Transformer*) [2] y BERT (*Bidirectional Encoder Representations from Transformers*) [3], 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 profesores como a estudiantes. Estos modelos están basados en lo que se conoce como inteligencia artificial generativa (IA generativa) [4] 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*) [5] 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 [6]. 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 [7-8], al que complementaremos con una base de datos vectorial Chroma [9] 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.

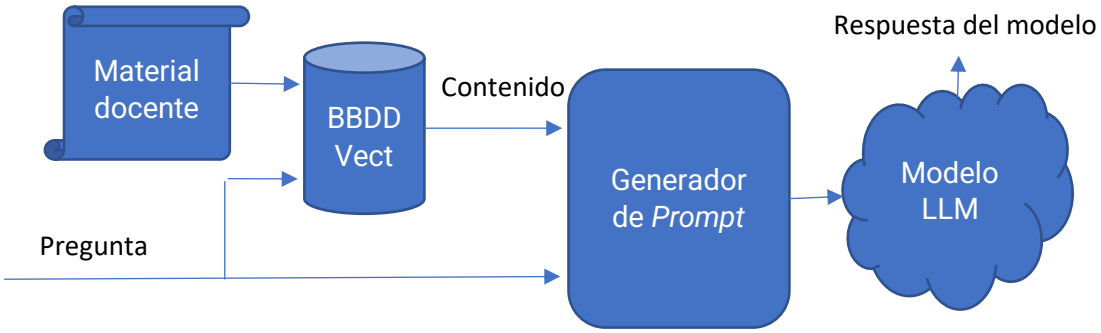

Figura 1. Arquitectura RAG.

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 profesor 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 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 realizar las preguntas al RAG de forma automática para evaluar los resultados, en vez de emplear el *chatbot*.

5. 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* (instruccies que se proporcionan a la IA) se usa únicamente la pregunta, que puede incluir las posibles respuestas en caso de ser multi-opción, 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 le parece o no correcta. En estos casos suele apostar por la pregunta 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 profesor valide la explicación para evitar errores en las respuestas al estudiantado.

5.1 Análisis de Estados Financieros

En la asignatura de 'Análisis de Estados Financieros' se han utilizado preguntas de respuesta múltiple con tres alternativas posibles. Las preguntas se han obtenido del banco de preguntas de preguntas en Moodle que se viene utilizando en la asignatura por el equipo docente de la asignatura 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.

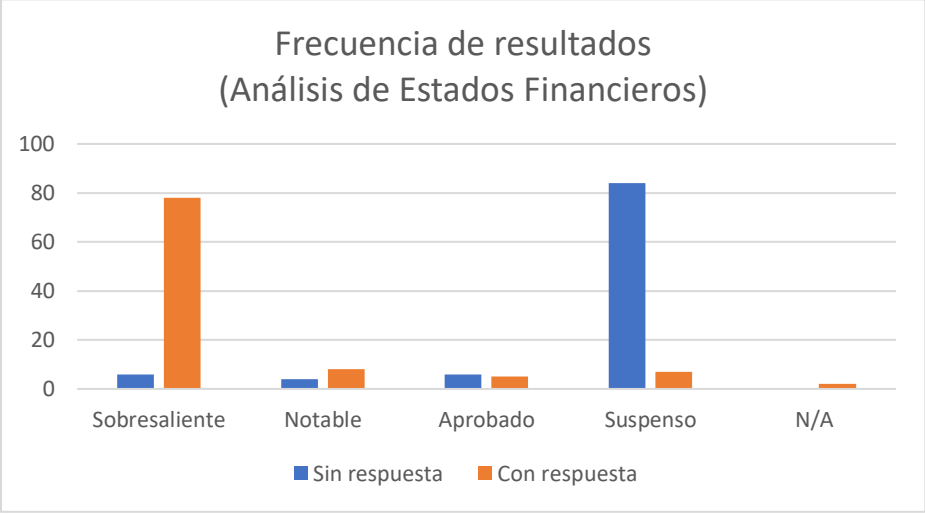

Figura 2. Resultados obtenidos con preguntas de Análisis de Estados Financieros

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 las 16 restantes, sólo 6 pueden considerarse completamente correctas. En general, se observa que el modelo responde mejor a preguntas generales, de conocimiento 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.

Cuando se le proporcionan las respuestas correctas al modelo, las proporciones se invierten. De las 100 preguntas analizadas, 78 fueron completamente correctas y tan sólo 7 fueron etiquetadas como 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 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).

5.2 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 concretitud, 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 los profesores de la asignatura y que sirve de rúbrica para la evaluación homogénea de las respuestas de los estudiantes cuando la corrección se realiza por varios profesores. En una primera prueba se realizarán 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 profesores expertos.

5.3 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 [10].

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.

Según se muestra en la figura 3, como ya ocurría en casos previos, resulta más útil el que el modelo conozca cuál es la respuesta correcta, independientemente de que conozca las cuatro opciones, para dirigir el contenido de la respuesta. El modelo en este caso suele funcionar mejor, con menos preguntas no válidas y un mayor porcentaje de respuestas completamente satisfactorias. Las preguntas con respuestas incorrectas suelen producirse cuando se refieren a la resolución de una pregunta de carácter numérico. En estos casos (5 de las 9 incorrectas), aunque trata de explicar el método para resolver el problema, no es capaz de operar correctamente los resultados, inventándose los valores intermedios y finales para alcanzar la solución. Hay que notar que 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 un estudiante promedio, lo que permitiría validar preguntas de un examen antes de ponérselo a los estudiantes. Por el contrario, se desaconseja su uso para resolución de dudas si no existe una supervisión previa por parte del profesor antes de proporcionar la respuesta.

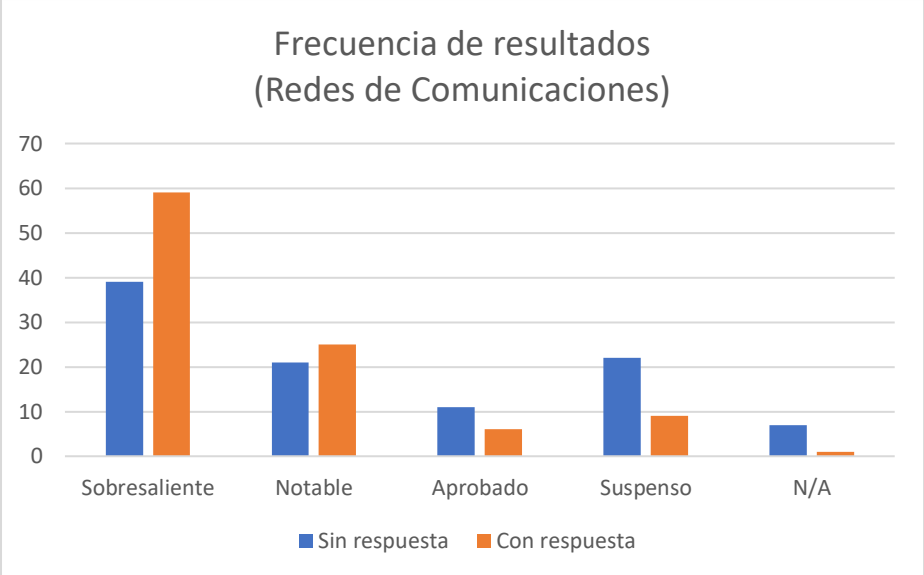

Figura 3. Resultados obtenidos con preguntas de Redes de Comunicaciones

6. Conclusiones

El uso de modelos LLM se ha extendido ampliamente entre la comunidad universitaria (p.e. *ChatGPT*). En este escenario debe ser el docente o el estudiante 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 de los estudiantes 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 a los estudiantes.

Según los resultados obtenidos, la actual versión debería estar en muchos casos supervisada por un profesor antes de entregar las respuestas para evitar que el modelo responda incorrectamente. No obstante, puede resultar útil para comprobar lo que un estudiante promedio respondería en un examen, y determinar de esta manera la dificultad del examen antes de que el estudiantado lo realice.

7. Referencias

[1] E. Mollick. Co-Intelligence. Living and Working with AI, 2024, ed. Portfolio Penguin
[2] G. Chemmalar Selvi. GPT (Generative Pre-Trained Transformer) - A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions, 12: 54608- 54649, 2024.
[3] J. Devlin, M.W. Chang, K. Lee and K. Toutanova. BERT: Pre- training of Deep Bidirectional Transformers for Language Understanding, pages 4171–4186, 2019.
[4] S. Feuerriegel, J. Hartmann, C. Janiesch and P. Zschech. Generative AI, Bus Inf Syst Eng 66(1):111–126, 2024.
[5] E. Sefika, P. Adrian. Retrieval-Augmented Generation-based Relation Extraction, 2024.
[6] Z. Bao, L. Liao-Liao, Z. Wu, Y. Zhou, D. Fan, M. Aibin, Y. Coady, and A. Brownsword. Delta Tensor: Efficient Vector and Tensor Storage in Delta Lake, 2024.
[7] Mistral AI, Mixtral of experts, https://mistral.ai/news/mixtral-of-experts/
[8] Huggingface, Mixtral-8x7B-Instruct-v0.1, https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1
[9] Chroma, The AI-native open-source embedding database, https://www.trychroma.com/
[10] J.A. Hernández, J. Conde, P. Reviriego, and G. Martínez Ruiz de Arcaute. Is ChatGPT capable of solving classical Communications and Networking problems?. TechRxiv. July 25, 2023. DOI: 10.36227/techrxiv.23727174.v1