XX Simposio CEA de Control Inteligente
25-27 de junio de 2025, Huelva

Aplicación de Modelos de Lenguaje de Gran Tamaño para la Recuperación de Información Legal

Martín, C.A.[a], Aguilar, R.M.[a] , Torres, J.M.[a], Bacallado M.A.[a], Alayón, S.[a]*, Savirón, G.E.[b]

[a] *Departamento de Ingeniería Informática y de Sistemas, Universidad de La Laguna, Camino San Francisco de Paula, 19, 38200, España*
[b] *Audiencia Provincial de Las Palmas de Gran Canaria, Sección 4 (Juzgado de lo Mercantil), España.*

To cite this article: Martín Galán, C.A., Aguilar Chinea, R.M., Torres Jorge, J.M., Bacallado López, M.A., Alayón Miranda, S., Savirón Díez, G.E., 2025. Aplicación de Modelos de Lenguaje de Gran Tamaño para la recuperación de información legal. XX Simposio CEA de Control Inteligente. https://doi.org/pendiente

Resumen

Este trabajo explora el uso de Modelos de Lenguaje de Gran Tamaño (LLMs), concretamente GPT-4o de OpenAI, para asistir en tareas de recuperación de información legal en el contexto del reto *Artificial Intelligence for Legal Assistance* (AILA). Se desarrollaron dos experimentos que evalúan la capacidad del modelo para identificar estatutos jurídicos relevantes ante consultas legales. En el primer experimento, se plantearon consultas individuales que relacionaban de forma directa cada estatuto con una consulta, obteniendo resultados limitados (precisión media del 18 %). En el segundo experimento, se amplió el contexto al presentar al modelo un conjunto mixto de estatutos, solicitando la identificación de los más relevantes. Esta estrategia logró una mejora significativa, alcanzando una tasa de recuperación de relevantes del 72,46 %. Los resultados evidencian que el modelo responde mejor cuando se le proporciona un entorno contextual más rico, lo cual respalda la viabilidad de incorporar en trabajos futuros un sistema de tipo Retrieval-Augmented Generation (RAG). Este permitiría prefiltrar el corpus legal en función de la similitud semántica, optimizando así la precisión y eficiencia del sistema de asistencia jurídica basado en IA.

Palabras clave: IA Generativa, LLM, Retrieval-Augmented Generation (RAG).

Application of Large Language Models for legal information retrieval

Abstract

This study explores the use of large language models (LLMs), specifically OpenAI's GPT-4o, to assist with legal information retrieval tasks within the framework of the *Artificial Intelligence for Legal Assistance* (AILA) challenge. Two experiments were conducted to assess the model's ability to identify relevant legal statutes in response to legal queries. In the first experiment, individual queries were directly matched to specific statutes, yielding limited results (average precision of 18%). In the second experiment, the context was expanded by presenting the model with a mixed set of statutes and requesting identification of the most relevant ones. This strategy led to a substantial improvement, achieving a relevant retrieval rate of 72.46%. The results indicate that the model performs better when provided with a richer contextual environment, supporting the feasibility of incorporating a Retrieval-Augmented Generation (RAG) system in future work. Such a system would enable pre-filtering of the legal corpus based on semantic similarity, thereby optimizing the accuracy and efficiency of AI-based legal assistance systems.

Keywords: Generative AI, LLM, Retrieval-Augmented Generation (RAG).

1. Introducción

En los sistemas jurídicos, el derecho emana principalmente de dos fuentes fundamentales: las leyes escritas y los precedentes judiciales. Las leyes codifican los principios legales aplicables a diversas situaciones, mientras que los precedentes (decisiones de casos previos) orientan a los juristas sobre cómo los tribunales han resuelto asuntos análogos en el pasado. Ambos pilares –normativa y jurisprudencia– son cruciales para la interpretación y aplicación del derecho, ya que permiten adaptar las reglas

*Autor para correspondencia: salayon@ull.edu.es

generales a casos concretos mediante el estudio de experiencias previas.

Con este trasfondo, surge el reto *Artificial Intelligence for Legal Assistance* (AILA, 2019), enfocado en la aplicación de la inteligencia artificial al dominio legal. El objetivo principal de AILA es desarrollar herramientas inteligentes capaces de identificar automáticamente casos previos y normativas relevantes que se apliquen a una situación jurídica dada. La tarea de vincular un problema legal nuevo con las leyes y precedentes pertinentes suele requerir extensas horas de investigación; por ello, contar con sistemas automatizados que agilicen este proceso resulta de gran importancia práctica para abogados y operadores jurídicos. En esencia, AILA busca automatizar y mejorar la investigación legal, facilitando que ante una determinada consulta se obtenga rápidamente un conjunto de documentos jurídicos relevantes (fallos judiciales y disposiciones legales aplicables) que sirvan de base para el análisis del caso.

Para alcanzar este objetivo, el reto plantea dos tareas principales. La Tarea 1 consiste en recuperar los casos judiciales más relevantes (precedentes jurisprudenciales) para una situación o consulta dada, a partir de un corpus de aproximadamente 3.000 documentos que recopilan sentencias de la Corte Suprema de la India. La Tarea 2 implica identificar cuáles son las leyes o estatutos legales más pertinentes relacionados con esa misma situación, eligiéndolos de un conjunto de 197 disposiciones normativas proporcionadas (cada una acompañada de su título oficial y descripción). En otras palabras, ante cada pregunta o asunto jurídico planteado, el sistema debe sugerir tanto jurisprudencia relevante (casos análogos ya resueltos) como legislación aplicable (artículos o secciones legales apropiadas) que podrían fundamentar el análisis de dicha cuestión.

El desarrollo de la competición estuvo sujeta a estrictas normas que buscaban garantizar la equidad y el enfoque en los datos proporcionados. Para mantener la objetividad del reto, no se permitía el uso de recursos externos relacionados con el dominio legal indio. En concreto, los participantes tenían prohibido apoyarse en documentos jurídicos ajenos al conjunto de datos proporcionado o emplear motores de búsqueda y sistemas legales preexistentes de la web para localizar casos o leyes pertinentes, de modo que las soluciones se basaran exclusivamente en la información dada por la organización del reto (Gain et al., 2021, Lalitha et al., 2022).

La evaluación de las soluciones propuestas se llevó a cabo mediante métricas estándar de rendimiento utilizadas en recuperación de información. En particular, se midieron la Precisión y el Recall (Cobertura), así como la Precisión Media – Mean Average Precision o MAP– y el F-score, entre otras. Estas métricas permitieron cuantificar de forma objetiva la eficacia de cada modelo al recuperar documentos relevantes. Por ejemplo, una mayor Precisión indica que el sistema devuelve pocos falsos positivos, y un mayor Recall refleja que recupera la mayoría de los documentos relevantes existentes. La métrica MAP resume la precisión a diferentes niveles de recall y es especialmente útil para evaluar rankings de documentos, mientras que el F-score ofrece una medida equilibrada que combina Precisión y Recall. Conforme a los criterios del reto, se valoraron positivamente aquellas soluciones capaces de identificar el mayor número de casos o leyes verdaderamente relevantes para cada consulta, minimizando al mismo tiempo la inclusión de resultados no pertinentes. En suma, el rendimiento de cada modelo se juzgó por su capacidad para hallar información jurídica relevante con exactitud y exhaustividad, comparando los resultados obtenidos con un conjunto de referencias establecido por los organizadores.

Finalmente, conviene destacar la relevancia de este reto en el contexto actual, donde la inteligencia artificial está teniendo un impacto creciente en el sector jurídico. La aplicación de técnicas de IA al derecho ofrece grandes oportunidades para mejorar la eficiencia en los procesos legales y facilitar el acceso a la justicia. Herramientas automatizadas como las promovidas por AILA pueden agilizar notablemente la investigación jurídica al analizar rápidamente volúmenes masivos de información legislativa y jurisprudencial, ahorrando tiempo a abogados y jueces en la búsqueda de precedentes y normativas aplicables. Asimismo, un sistema capaz de sugerir casos y leyes relevantes brinda una orientación valiosa a los ciudadanos, incluso antes de que recurran a un abogado para solicitar ayuda. De este modo, iniciativas como AILA no solo aumentan la eficiencia y rigor en la preparación de casos legales, sino que también contribuyen a democratizar el conocimiento legal, haciendo más asequible y rápida la identificación de normas y precedentes pertinentes para cualquier persona que enfrente una duda jurídica.

2. Materiales y Métodos

Este trabajo desarrolla una estrategia para resolver las tareas planteadas en el reto mencionado en apartados anteriores aprovechando las capacidades de los modelos LLMs sin tener que recurrir a otras técnicas de Machine Learning (ML) o Aprendizaje Profundo (DL). Analizando los resultados de los diferentes experimentos realizados en su desarrollo se podrá determinar una estrategia para resolver el reto planteado utilizando las ventajas que supone disponer de modelos LLMs entrenados previamente con grandes volúmenes de información.

Para llevar a cabo los experimentos del presente trabajo, se desarrollaron diferentes programas utilizando el lenguaje de programación Python (Python, 2025). El objetivo principal de estos programas fue estructurar y organizar de manera eficiente los datos suministrados en el marco del reto *Artificial Intelligence for Legal Assistance* (AILA), facilitando así su posterior procesamiento.

Los datos proporcionados por el reto comprenden tres tipos principales de documentos: estatutos legales, casos judiciales y consultas jurídicas. Con el fin de permitir un acceso eficiente a cada uno de estos componentes, se construyeron tres diccionarios de datos independientes: uno para los estatutos, otro para los casos, y un tercero para las consultas.

El diccionario de estatutos legales (*s_dict*) fue construido a partir de archivos doscientos ficheros individuales, donde cada documento representa una ley o sección específica identificada con un prefijo "S" seguido de un número (por ejemplo, S1.txt). En este diccionario, las claves corresponden al identificador del estatuto (sin extensión), mientras que los valores almacenan el contenido textual completo del documento legal.

Esta estructura permite una recuperación rápida de los estatutos relevantes durante el análisis.

El diccionario de casos judiciales (*c_dict*) fue generado a partir de 2914 documentos, los cuales almacenan transcripciones de sentencias de la Corte Suprema de la India, nombradas bajo el patrón *C<id>.txt*. Al igual que en el caso anterior, se asignaron como claves los identificadores de los casos y como valores, los textos completos de las sentencias.

El diccionario de consultas (*query_dict*), se crea a partir de un único archivo que contiene una serie de cuestiones legales, cada una identificada con un código único (por ejemplo, AILA_Q1) seguido del texto de la consulta, separados por el delimitador "||". La organización de estas preguntas en un diccionario facilita su tratamiento individual y su vinculación con los documentos legales y jurisprudenciales relevantes.

Con el fin de validar los resultados obtenidos por los modelos y medir su rendimiento en términos de recuperación de información relevante, se implementaron estructuras adicionales en Python destinadas a almacenar las relaciones de relevancia entre las consultas jurídicas y los documentos legales correspondientes. En particular, se construyeron dos diccionarios: uno para asociar cada consulta con los estatutos legales relevantes *(s_relevance_dict*), y otro para relacionar cada consulta con los casos judiciales relevantes (*c_relevance_dict*). Ambos diccionarios fueron generados a partir de archivos de referencia que indican, para cada consulta, qué documentos deben considerarse como relevantes.

Cada línea del archivo de solución contiene referencia a una consulta, a un documento (estatuto o caso), y una etiqueta binaria de relevancia. En el caso de que esta etiqueta sea "1", el documento es considerado relevante para dicha consulta. Así, los diccionarios se construyen utilizando como clave el identificador de la consulta (por ejemplo, AILA_Q1) y como valor una lista con los identificadores de los estatutos o casos asociados. Estas estructuras de relevancia constituyen una parte esencial del marco experimental, ya que permiten evaluar de forma precisa la capacidad del sistema para recuperar los documentos jurídicos pertinentes ante una determinada consulta, y sirven como base para calcular métricas de evaluación de los diferentes experimentos.

Para la realización de los experimentos y el procesamiento automatizado de las consultas jurídicas, se utilizó el modelo de lenguaje GPT-4o de OpenAI, accediendo a sus capacidades mediante la API oficial de OpenAI. Este modelo ofrece mejoras sustanciales en tareas de comprensión y generación de lenguaje natural, lo que lo convierte en una herramienta idónea para contextos jurídicos donde la interpretación precisa de los textos es fundamental. A través de scripts desarrollados en Python, se enviaron de forma programática las consultas estructuradas previamente, recibiendo como respuestas las leyes o casos que eran considerados relevantes para cada consulta. Este enfoque permitió automatizar los experimentos y analizar de forma sistemática los resultados generados por el sistema.

La integración con la API de OpenAI se realizó conforme a la documentación oficial disponible en (API de OpenAI, 2025). La automatización del proceso permitió ejecutar múltiples consultas de manera eficiente, lo que resulta especialmente útil para la evaluación comparativa de resultados y el desarrollo iterativo del sistema.

3. Metodología propuesta

Con el objetivo de evaluar la capacidad de los modelos de lenguaje de gran tamaño (LLMs), específicamente GPT-4o de OpenAI (GPT-4o, 2025), para asistir en tareas de análisis jurídico, se diseñaron una serie de experimentos alineados con los desafíos planteados en el reto *Artificial Intelligence for Legal Assistance* (AILA). Estos experimentos tienen como propósito analizar la calidad de las respuestas generadas por el modelo en relación con la identificación de documentos legales relevantes —estatutos y casos judiciales— para distintas consultas jurídicas.

La metodología seguida se basa en la automatización del diálogo con el modelo mediante el uso de su API, a fin de garantizar un entorno de pruebas replicable y escalable. Cada experimento explora una estrategia diferente de interacción con el modelo, evaluando su rendimiento en la tarea de recuperación de información legal relevante. En todos los casos, se utilizaron los diccionarios previamente construidos (*s_dict, c_dict, query_dict*) como punto de partida para la generación sistemática de las consultas.

3.1. Primer experimento: Evaluación directa de relevancia entre estatuto y consulta.

El primer experimento consistió en una estrategia de evaluación uno a uno, mediante la cual se analizó si el modelo es capaz de determinar, de manera explícita, la relevancia de un estatuto legal con respecto a una consulta jurídica determinada. Para ello, se recorrieron todas las combinaciones posibles entre las consultas del diccionario *query_dict* y los estatutos del diccionario *s_dict*. En cada caso, se formuló una consulta directa al modelo preguntando si un determinado estatuto resultaba relevante para una situación descrita en una consulta específica.

La interacción con el modelo se llevó a cabo mediante el siguiente prompt estructurado:

"*Given the following information about statutes from Indian law, please indicate if the indicated statute is relevant for the described situation that had led to filing a case in an Indian court of law, answering with a yes or no. Statutes are considered relevant to a situation if they discuss a situation similar to that in the query, as judged by law experts.*
Statute: {s_dict[s]}
Situation: {query_dict[q]}"

Este prompt fue diseñado para presentar al modelo tanto el contenido del estatuto como la descripción de la situación legal (consulta), solicitando una respuesta categórica ("yes" o "no"). La simplicidad de la respuesta esperada permite una evaluación directa de la precisión del modelo, comparando sus respuestas con los datos de referencia almacenados en *s_relevance_dict*.

Este enfoque permite examinar de manera granular la capacidad del modelo para reconocer relaciones semánticas y contextuales entre textos jurídicos, simulando una tarea de clasificación binaria basada en criterios de similitud legal. La

sistematización del experimento a través de código Python permitió aplicar esta metodología a todo el conjunto de datos, generando una base de resultados.

3.2. Segundo experimento: Evaluación del modelo en contexto ampliado

En el caso de que la evaluación basada en consultas individuales (uno a uno) no ofreciera resultados satisfactorios para el problema en estudio, se planteó un segundo experimento con el objetivo de analizar si el rendimiento del modelo puede mejorar al proporcionarle un mayor contexto legal. Esta línea de trabajo responde a un interés particular en evaluar la viabilidad de incorporar en desarrollos futuros un módulo de tipo Retrieval-Augmented Generation (RAG) que permita, de forma automatizada, extraer un subconjunto de documentos jurídicos semánticamente cercanos a una consulta determinada, como paso previo a la interacción con el modelo de lenguaje.

En este experimento, se recorren todas las consultas del diccionario *query_dict*, y para cada una se construye un prompt enriquecido en el que se presentan al modelo múltiples estatutos al mismo tiempo. A diferencia del primer experimento, donde se analizaban individualmente, aquí el modelo debe seleccionar las leyes más relevantes desde un conjunto más amplio. Para garantizar que este conjunto incluye tanto estatutos pertinentes como no pertinentes, se diseña una estrategia controlada: para cada consulta, se incluye el número exacto de estatutos definidos como relevantes en la solución del reto (N), y se añaden aleatoriamente (sin reemplazo) N × factor estatutos en total, donde factor es un parámetro configurable (inicialmente fijado en 4).

Por ejemplo, si una consulta del reto (como AILA_Q41) tiene 5 estatutos relevantes según la solución de referencia, se construye un conjunto de 20 estatutos (5 × 4), de los cuales 5 son efectivamente relevantes y 15 se seleccionan al azar del resto del corpus. Este conjunto se presenta al modelo acompañado de la descripción de la situación jurídica correspondiente, y se le solicita que indique cuáles son los N estatutos más relevantes del total mostrado.

El prompt utilizado en este experimento sigue la siguiente estructura:

" *Given the following information about statutes from Indian law, please indicate which ones are the {n_relevant} most relevant for the described situation that had led to filing a case in an Indian court of law. Statutes are considered relevant to a situation if they discusses a situation similar to that in the query, as judged by law experts.*

Statutes: {ask_for_query}

Situation: {query_dict[query]}"

El texto *ask_for_query* representa la concatenación de los textos de los estatutos seleccionados para cada consulta, y *query_dict[query]* corresponde a la situación legal en cuestión.

Los dos experimentos planteados ofrecen aproximaciones complementarias para evaluar la capacidad del modelo LLM (en este caso, GPT-4o) de identificar estatutos jurídicos relevantes a partir de una consulta legal. Con las consultas individuales se ofrece una evaluación exhaustiva y detallada, pero tiene algunas limitaciones:

a) No considera el contexto comparativo entre estatutos, lo que puede limitar la capacidad del modelo para priorizar.
b) Es menos realista desde el punto de vista del uso en sistemas de recuperación, donde se espera trabajar con subconjuntos prefiltrados.

El segundo experimento de selección en contexto amplio plantea un entorno más cercano al de una aplicación real: se proporciona al modelo un subconjunto de y se le solicita seleccionar los más pertinentes. Esta estrategia refleja mejor un escenario en el que la IA debe filtrar y priorizar leyes a partir de información contextual. De esta forma se permite evaluar la capacidad del modelo de distinguir y seleccionar dentro de un conjunto razonablemente grande y sirve como base para determinar si la incorporación de un módulo de tipo Retrieval-Augmented Generation (RAG) es viable y beneficiosa.

En este sentido, si los resultados del segundo experimento muestran que el modelo es capaz de identificar consistentemente la mayoría de los estatutos relevantes dentro de conjuntos mixtos, se justifica con más fuerza la idea de ampliar el proyecto hacia un sistema RAG. Dicho sistema podría encargarse de reducir el conjunto inicial de estatutos a partir de criterios semánticos, y luego dejar al modelo LLM la tarea de priorizar entre los más similares, optimizando así tanto eficiencia como precisión.

4. Conclusiones

4.1. Resultados del primer experimento: Evaluación directa de relevancia entre estatuto y consulta.

Para medir la calidad de las respuestas generadas por el modelo GPT-4o en relación con la relevancia de los estatutos frente a cada consulta, se utilizaron métricas estándar de evaluación en tareas de clasificación binaria: precisión (*precision*), recuperación (*recall*) y F1-score. Estas métricas permiten cuantificar, de forma objetiva, el grado de acierto del modelo en la identificación de leyes pertinentes según los datos de referencia.

En el contexto del experimento, para cada consulta (query), se genera una predicción binaria para cada ley (statute), indicando si el modelo la considera relevante o no relevante. Esta predicción se compara con la solución esperada, permitiendo clasificar los resultados en cuatro categorías:

- Aciertos positivos (TP - true positives): leyes que el modelo consideró relevantes y que efectivamente lo eran según la solución.
- Aciertos negativos (TN - true negatives): leyes que el modelo consideró no relevantes y que en efecto no lo eran.
- Falsos positivos (FP): leyes que el modelo consideró relevantes, pero que en realidad no lo eran.
- Falsos negativos (FN): leyes que el modelo consideró no relevantes, aunque sí lo eran según la solución.

A partir de estas cantidades, se definen las métricas de evaluación del siguiente modo (Chauhan, 2023):

- Precisión: mide la proporción de leyes consideradas relevantes por el modelo que realmente lo eran. Se calcula como precisión = TP / (TP + FP)
- Recall (sensibilidad o recuperación): indica la proporción de leyes relevantes que el modelo fue capaz de identificar correctamente. Se calcula como recall = TP / (TP + FN)

- F1-score es la media armónica entre la precisión y el recall, y proporciona una medida balanceada que penaliza tanto los falsos positivos como los falsos negativos. Se calcula según la fórmula F1 = 2 * precisión * recall / (precisión + recall)

Durante la ejecución del experimento se observaron que algunos problemas de respuesta al modelo. Algunos de estos problemas eran producidos por fallos en el conjunto de datos (consultas que daban como relevantes leyes que no existían) y otros por motivos de la consulta al modelo en lo que respondía con deducciones y análisis de sentimientos en vez de clasificar de forma binaria.

En los resultados se observó que acierta significativamente menor la no relevancia de una ley en una consulta mientras que muestra un resultado pobre averiguando aquellas que son relevantes. La precisión media obtenida fue de 0,18. Esta baja precisión justifica la búsqueda de otras soluciones y por tanto la realización del segundo experimento para poder comprobar si aumentando el contexto en la consulta, y por lo tanto suministrando más información para predecir, mejora el resultado.

4.2. Resultados del segundo experimento: Evaluación del modelo en contexto ampliado.

En el segundo experimento, la evaluación no se basa en métricas tradicionales como la precisión o el F1-score, ya que no se trata de una tarea de clasificación binaria. En su lugar, se propone una métrica específica, que aquí sugerimos denominar tasa de recuperación de relevantes (TRR) y que definimos como TRR = nRS * 100 / N

Siendo:

- nRS el número de leyes relevantes seleccionadas del total de N relevantes que el reto daba como solución.
- N es el número total de leyes que el modelo debía seleccionar (es decir, el número de relevantes definidos por la solución para esa consulta).

Por ejemplo, si para una consulta el modelo debía seleccionar 5 leyes y logró incluir 4 de los que eran realmente relevantes, la TRR sería del 80%. Esto indica una muy buena capacidad de priorización del modelo, especialmente si el conjunto total de entrada incluía muchos distractores.

El número de distractores se seleccionaba en el experimento con un factor (en este caso se fijó el factor en 4), lo que indicaba que, si para una consulta había 5 leyes relevantes, el conjunto total de leyes entre las que tenía que escoger era 20.

La métrica TRR nos permite comparar el rendimiento del modelo en distintas consultas y resulta intuitiva, ya que una TRR del 100% significa que el modelo no omitió ninguna ley relevante en su respuesta. Lo más importante en nuestro caso de estudio es que puede dar una pista clara sobre la viabilidad de aplicar estrategias de prefiltrado, de modo que si el modelo logra altos valores de TRR con conjuntos limitados (como 4×N), puede confiarse en su capacidad de discriminación en entornos RAG.

Obtuvimos que la tasa media de recuperación de relevantes es del 72,46%, lo cual indica que el modelo se comporta mejor cuando en la consulta incluimos mayor información de contexto.

Los experimentos realizados en este trabajo han permitido analizar el comportamiento del modelo de lenguaje GPT-4o de OpenAI en el contexto del reto *Artificial Intelligence for Legal Assistance* (AILA), evaluando su capacidad para identificar estatutos legales relevantes ante distintas situaciones jurídicas. Los resultados obtenidos evidencian que el rendimiento del modelo varía significativamente en función del diseño del experimento y del contexto informativo proporcionado.

En conclusión, en el primer experimento, en el que se planteaban consultas individuales sobre la relevancia de cada estatuto de forma aislada, el modelo presentó un rendimiento limitado, especialmente en lo que respecta a la identificación de leyes relevantes. La precisión media obtenida fue de tan solo 0,18, lo que revela una alta proporción de falsos positivos y una dificultad evidente para discriminar relevancia sin un marco comparativo. Estos resultados reflejan las limitaciones de abordar la tarea en un entorno excesivamente fragmentado, sin ofrecer al modelo suficiente contexto para apoyar sus decisiones.

En cambio, el segundo experimento, basado en la presentación conjunta de múltiples estatutos en una misma consulta y solicitando al modelo que identificara los más relevantes, mostró una mejora sustancial. La Tasa de Recuperación de Relevantes (TRR) alcanzó un valor medio del 72,46 %, lo que indica que el modelo fue capaz de incluir en sus respuestas una parte considerable de los estatutos correctos incluso dentro de conjuntos mixtos, con presencia de distractores. Esta mejora sugiere que, al contar con mayor información contextual, el modelo puede realizar comparaciones más efectivas y emitir juicios más acertados sobre la pertinencia jurídica de los documentos.

A partir de estos hallazgos, se desprende una conclusión clave: el rendimiento del modelo puede optimizarse significativamente si se le provee de un contexto semánticamente relevante previo a la generación de respuestas. En este sentido, se justifica plenamente la exploración de una línea futura de trabajo basada en la implementación de un sistema Retrieval-Augmented Generation (RAG). Este tipo de arquitectura permitiría prefiltrar, a partir de criterios de similitud semántica, un subconjunto reducido de estatutos jurídicos potencialmente relevantes para cada consulta, sobre el cual el modelo LLM podría aplicar sus capacidades de análisis y priorización (Lewis et al., 2020).

Así, un sistema RAG ofrecería una solución más eficiente, escalable y precisa para tareas complejas de asistencia legal automatizada, integrando mecanismos de recuperación de información con modelos generativos de alto rendimiento. El presente trabajo sienta las bases para dicha evolución, demostrando la relevancia del contexto en la toma de decisiones del modelo y ofreciendo una metodología sólida para su evaluación y futura expansión.

Agradecimientos

Este trabajo ha sido financiado por la Fundación CajaCanarias y la Fundación La Caixa [número de subvención 2023DIG11]

Referencias

AILA. Reto “Artificial Intelligence for Legal Assistance” (2019). Sitio web: https://sites.google.com/view/fire-2019-aila/. Último acceso: 9/06/2025.

API de OpenAI (2025). Documentación. Sitio web: https://platform.openai.com/docs/overview. Último acceso: 9/06/2025.

Chauhan N.S. Métricas de evaluación de modelos en el Aprendizaje Automático (2023). Disponible en: https://www.datasource.ai/es/data-science-articles/metricas-de-evaluacion-de-modelos-en-el-aprendizaje-automatico. Último acceso: 9/06/2025.

Gain, Baban, Dibyanayan Bandyopadhyay, Arkadipta De, Tanik Saikh and Asif Ekbal (2021). IITP at AILA 2019: System Report for Artificial Intelligence for Legal Assistance Shared Task. *Fire*.

GPT-4o, OpenAI (2025). Sitio web: https://openai.com/es-ES/index/hello-gpt-4o/. Último acceso: 9/06/2025.

Lalitha, Y. S., Raju, N. V. G., Teja, V. R., Sravani, P., Reddy, E. S. (2022). AI enabled legal assistance system: A case study. International Journal of Health Sciences, 6(S3), 6835–6844.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, *33*, 9459-9474.

Python. Sitio web: https://www.python.org/. Último acceso: 9/06/2025.