

# Mex-IA: Chatbot basado en RAG enfocado al apoyo legal del ciudadano

Daniel Lima López<sup>1</sup>, Héctor Anaya Sánchez<sup>1</sup>, Francisco David Camacho González<sup>1</sup>

<sup>1</sup>Coordinación de Ciencias Computacionales, Instituto Nacional de Astrofísica, Óptica y Electrónica

E-mails: limalopezdaniel2@gmail.com, hector.anaya@inaoep.mx, francisco.david.9.7@gmail.com

## 1. Generalidades

### 1.1. Generalidades del proyecto

El desconocimiento o la falta de información de los ciudadanos mexicanos sobre el contenido de la constitución representa un grave problema. Entre las principales consecuencias de esta situación se encuentran: la manipulación o estafa de individuos a través de estrategias de desinformación, confusión con respecto a los derechos y obligaciones de los mexicanos, y el abuso por parte de autoridades federales y estatales, etc. Si bien el contenido de la constitución mexicana se encuentra públicamente disponible en diversos medios de información, la redacción o el desconocimiento de términos legales dificulta el entendimiento de su contenido al ciudadano promedio, a tal grado que una gran variedad de trámites institucionales básicos o simples consultas legales requieren del apoyo de profesionales del derecho. Servicio al cual gran parte de la población no tiene acceso por limitaciones económicas. Este problema también representa un gran desafío en el caso de migrantes, los cuales constantemente se enfrentan a diversos abusos debido al desconocimiento de sus derechos en nuestro país.

Este proyecto propone implementar un sistema de Chatbot enfocado a responder consultas de carácter legal, utilizando la información contenida en el documento oficial de la constitución mexicana. La herramienta emplea la técnica RAG (Retrieval Augmented Generation), que, dada una consulta del usuario, identifica los documentos que guardan más relación con la misma. En este caso, cada documento corresponde a un artículo de la constitución. Una vez que se han identificado los artículos que guardan una mayor relación con el contexto de la consulta del usuario, se recurre a un modelo de lenguaje, encargado de dar respuesta a la consulta considerando la información identificada. Además, se implementa la técnica one-shot learning para entrenar al modelo de lenguaje, permitiéndole responder a las consultas indicando las fuentes consideradas para generar la respuesta final. El diagrama general del sistema implementado se muestra en la Figura 1.1.

El objetivo del proyecto es ofrecer una herramienta que facilite la consulta y comprensión de las leyes mexicanas. Esta herramienta no pretende sustituir la asesoría de un profesional, sino proporcionar un primer acercamiento al usuario promedio para ayudar en la obtención de información legal. El sistema se basa en los artículos de la constitución mexicana, pero el proyecto puede expandirse a una amplia variedad de reglamentos o documentos oficiales, como el reglamento de tránsito o los requisitos para trámites oficiales. Además, el sistema puede integrarse como un asistente virtual en diversos portales oficiales, sirviendo como apoyo

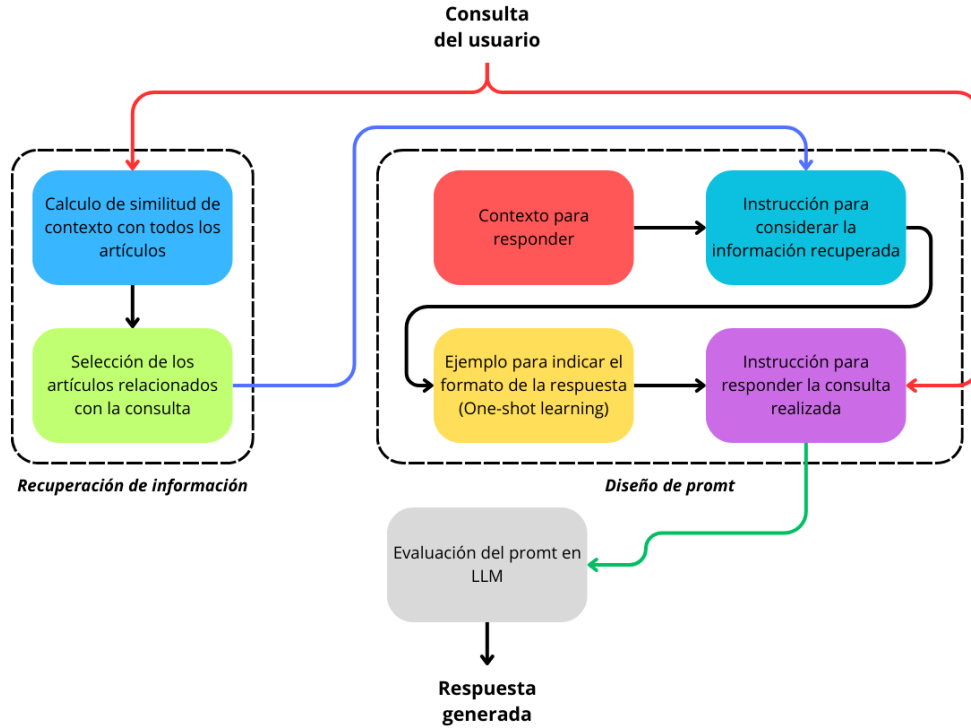


Figura 1: Diseño del modelo implementado.

adicional para aquellos que necesiten realizar consultas sobre procesos legales o resolver dudas sobre el contenido de la constitución.

## 1.2. Oportunidad Detectada

Este proyecto satisface una necesidad clara: el acceso rápido y comprensible a información legal, especialmente en temas relacionados con la Constitución. Actualmente, muchos ciudadanos no conocen sus derechos o carecen de los recursos necesarios para acceder a asesoramiento legal. Mex-IA se presenta como una solución a esta problemática, facilitando el acceso a información legal sin necesidad de recurrir a abogados o servicios costosos, que suelen ser inaccesibles para la mayor parte de la población. Además, simplifica la interpretación de un lenguaje legal que puede ser complejo y difícil de entender para quienes no tienen formación en derecho.

Estudios que muestran el porcentaje de la población en México que no tiene acceso fácil a servicios legales. Un informe del World Justice Project reveló que casi la mitad de los mexicanos (49%) ha enfrentado al menos un problema legal en los últimos dos años, pero solo el 23% de ellos buscó asistencia legal formal [3]. Asimismo, las disputas relacionadas con el consumo, la vivienda y el trabajo son los problemas más frecuentes entre los mexicanos. Sin embargo, el acceso a servicios legales formales sigue siendo restringido para gran parte de la población. Esto sugiere una necesidad crítica de democratizar y mejorar el acceso a la justicia a través de herramientas tecnológicas y asesoría automatizada, como lo que se busca en el proyecto Mex-IA. Asimismo, la creciente adopción de la tecnología en la vida diaria de la población para resolver problemas cotidianos refuerza la noción de que un chatbot legal puede ser una herramienta útil para disminuir la brecha en el acceso a la información jurídica.

El propósito del proyecto Mex-IA es claro: ofrecer una herramienta automatizada y accesible que permita a los ciudadanos conocer sus derechos constitucionales de manera clara y directa, mejorando así su capacidad para tomar decisiones informadas sobre temas legales. El modelo propuesto busca democratizar el acceso a la información legal, asegurando que cualquier

persona, independientemente de su nivel educativo o económico, pueda entender y ejercer sus derechos de forma adecuada.

### 1.3. Generalidades Técnicas

El proyecto Mex-IA utiliza una serie de componentes tecnológicos que incluyen hardware, software, materiales, y procesos industriales. El software empleado incluye la API de OpenAI GPT-4 para generar respuestas en lenguaje natural, así como la librería LlamaIndex para facilitar la implementación de sistemas RAG [2]. Se utilizan herramientas como Tesseract OCR para extraer texto de documentos en formato PDF cuando es necesario [1]. Todas estas herramientas se implementaron utilizando Python como el lenguaje principal. No se utiliza firmware específico, ya que el proyecto no depende de hardware embebido, pero sí se cuenta con la base de datos digital de la Constitución de México como material esencial, además de los modelos preentrenados de OpenAI.

La solución teórica propuesta es un chatbot basado en RAG (Retrieval-Augmented Generation) que permite a los usuarios consultar la Constitución de México y recibir respuestas claras basadas en su contenido. Este sistema podría convertirse en un producto tangible, ya sea en forma de una aplicación móvil o una interfaz web, donde los ciudadanos interactuarían en tiempo real para recibir asesoría legal. La interacción es sencilla: el usuario formula una pregunta, el sistema consulta la base de datos de la Constitución, y la respuesta se genera mediante GPT-4o-mini, citando los artículos relevantes.

En cuanto al desarrollo del prototipo (PoC), se utilizó la metodología Scrum. Durante el sprint planning, se definen tareas específicas como la implementación de la generación de lenguaje natural y la integración de la base de datos de la Constitución. Se realizaron reuniones diarias (daily standups) para asegurar la colaboración y el avance continuo del proyecto, al final de cada sprint se lleva a cabo una evaluación de resultados para identificar mejoras. El flujo de trabajo sigue una estructura clara: el usuario ingresa una consulta, el sistema busca los artículos relevantes, y se genera una respuesta utilizando GPT-4. Esto garantiza que el producto final sea iterativo y evolucione constantemente según el feedback recibido.

El proyecto se encuentra actualmente en un nivel de TRL 4 (Technology Readiness Level), lo que significa que se ha validado la tecnología en un entorno de laboratorio y las pruebas han demostrado que el chatbot puede generar respuestas adecuadas basadas en preguntas legales. Para avanzar al TRL 5, los siguientes pasos incluyen la implementación de una versión beta en un entorno real, como una página web pública donde los usuarios puedan interactuar con el chatbot, así como la evaluación de las respuestas. En el TRL 6, se optimizará el rendimiento, mejorando el tiempo de respuesta y la capacidad de procesar múltiples consultas simultáneamente. Finalmente, en el TRL 7, se integrarán otras bases de datos legales para ampliar el alcance del chatbot, permitiendo que responda preguntas no solo sobre la Constitución de México, sino también sobre otras leyes y regulaciones aplicables.

Este enfoque asegura que el proyecto Mex-IA evolucione desde una prueba de concepto hasta una herramienta robusta que pueda ser implementada a nivel público, mejorando el acceso a la justicia para los ciudadanos de México.

El desarrollo del proyecto se realizó mediante las pruebas reportadas en la Tabla 1.

### 1.4. Innovación con IA

El impacto del uso de herramientas de inteligencia artificial en nuestro proyecto elimina la necesidad de realizar un entrenamiento complejo para implementar un Chatbot. Por consiguiente, gran parte de los esfuerzos se concentraron en optimizar calidad de las respuestas generadas y mejorar la etapa de búsqueda de información para generar cada respuesta.

En cuanto a la etapa de procesamiento de datos, se usaron técnicas como Expresiones Regulares (RE, Regular Expressions) y Reconocimiento Óptico de Caracteres (OCR, Optical Character Recognition) para extraer contenido de archivos PDF. La primera técnica se usó para

Nombre de la prueba	Tipo de prueba	Resultado-conclusiones
Prueba de integración de artículos	Simulación	La integración de los artículos de la Constitución en sistema RAG fue exitosa, permitiendo una recuperación eficiente del texto para pruebas internas.
Prueba de generación de respuestas	Simulación	Se verificó que las respuestas generadas por GPT-4 fueran coherentes y contextualizadas con base en los artículos de la Constitución recuperados.
Prueba de carga en laboratorio	Laboratorio	Se evaluaron 15 preguntas para verificar las respuestas generadas

Tabla 1: Pruebas realizadas durante el desarrollo del proyecto

extraer información de archivos que permiten la lectura como caracteres, mientras que la segunda técnica se usó para la transcripción de texto en imágenes. Las herramientas usadas para implementar el Chatbot son Modelos de Lenguaje (LLMs, Large Language Models) y Generación Aumentada por Recuperación (RAG, Retrieval Augmented Generation). La técnica RAG es usada para la identificación de la información más relevante para responder a las consultas de los usuarios, mientras que los modelos de lenguaje son usados para generar la respuesta final considerando la información identificada. Todas las herramientas fueron utilizadas a un nivel intermedio para facilitar el diseño de cada etapa del proyecto.

El modelo propuesto considera ciertos aspectos para garantizar el uso ético de herramientas de inteligencia artificial. En primer lugar, el modelo garantiza la privacidad del usuario, ya que recibe únicamente sus consultas, evitando la transferencia de información sensible. De igual manera, se evita que el modelo tenga sesgo, ya que las respuestas se generan sin considerar las características del usuario. Otro aspecto a considerar es la veracidad de las respuestas. Con respecto a esta situación, se optó por realizar un preprocesamiento exhaustivo del contenido de la constitución para garantizar que el proceso de búsqueda de información se realice sobre información validada. Además, el preprocesamiento de la información incluye el etiquetado de cada uno de los artículos de la constitución, de tal manera que el modelo, además de tener capacidad de consultar la información de cada artículo, también es capaz de especificar que artículos consulto para generar la respuesta, facilitando la evaluación del modelo y evitando la generación de respuestas espontáneas sin sustento de información.

### 1.5. Equipo de Desarrollo de la PoC

El equipo de desarrollo está compuesto por tres integrantes, todos egresados de la licenciatura en física de diferentes universidades. Actualmente, los integrantes han cubierto el plan de estudios de la maestría en ciencias computacionales. En cuanto a las tecnologías involucradas en el desarrollo del proyecto, cada integrante ha contribuido en más de una de las siguientes disciplinas:

- Extracción de información: La experiencia de los colaboradores en esta etapa fue crucial para acceder a fuentes oficiales y filtrar datos relevantes de manera eficiente, garantizando que la información utilizada sea precisa y confiable.
- RAG: Las habilidades de los colaboradores en esta fase permitieron la integración del proceso de búsqueda de información con la generación de texto. Lo cual resultó en la generación de respuestas coherentes y contextuales, mejorando la interacción con los usuarios.
- LLM: El conocimiento de los colaboradores en esta etapa permitió el ajuste adecuado de los hiperparámetros del modelo de lenguaje, garantizando la generación de respuestas coherentes.

## 2. Escenario de Pruebas según TRL

### 2.1. TRL 4 – 5: Prototipo a nivel Simulación

Medir la calidad de las respuestas proporcionadas por el modelo es de las tareas más complicadas a las que nos enfrentamos, pues no existe una métrica intuitiva que nos indique que tan acertadas son las respuestas generadas por el modelo implementado, y si estas citan a las fuentes correctas. Una solución sencilla es revisar las respuestas personalmente y verificar tanto la información como la fuente citada. Por lo tanto, se analizaron las respuestas generadas por el modelo al realizar 15 preguntas diferentes. Algunos ejemplos de estas preguntas incluyen: ¿Cuáles son mis derechos como trabajador?, ¿Cuál es la diferencia entre un delito y una falta cívica? Y ¿Qué es un delito doloso?

Por ejemplo, en cuanto a la pregunta ¿Cuáles son mis derechos como trabajador?, el modelo genera la siguiente respuesta:

*“Como trabajador, tienes derechos fundamentales que incluyen, entre otros:*

- *Derecho a un trabajo digno y socialmente útil.*
- *Jornada máxima de trabajo de ocho horas diarias y siete horas nocturnas.*
- *Derecho a un día de descanso por cada seis días de trabajo.*
- *Vacaciones de al menos veinte días al año.*
- *Salario mínimo que debe ser suficiente para satisfacer las necesidades normales de una familia.*
- *Igualdad de salario por trabajo igual, sin distinción de sexo.*
- *Derecho a la capacitación y adiestramiento.*
- *Protección contra despidos injustificados, con derecho a reinstalación o indemnización.*
- *Derecho a organizarse en sindicatos y a la huelga.*

*Fuente: Artículo 123 de la Constitución de México.”*

Note que la respuesta proporcionada enlista los principales derechos de los trabajadores y cita al artículo de la constitución que aborda los derechos laborales.

Por otro lado, la respuesta a la pregunta ¿Cuáles son los impuestos que debo pagar por tener una propiedad? Es la siguiente:

*“Los impuestos que debes pagar por tener una propiedad incluyen las contribuciones que establezcan los Estados sobre la propiedad inmobiliaria, así como las tasas adicionales que puedan aplicarse por su fraccionamiento, división, consolidación, traslación y mejora. Si no pagas estos impuestos, podrías enfrentar sanciones que van desde multas hasta la posibilidad de que se inicie un procedimiento de embargo sobre la propiedad.*

*Fuente: Artículo 115, incisos a) y b)”*

En este caso, el modelo produce una respuesta adecuada y menciona correctamente el artículo consultado. Además, observé que también identifica los incisos específicos de donde se está extrayendo la información.

Cabe destacar que en ambos casos el modelo logra generar una respuesta válida y que guarda relación con la información identificada por el sistema RAG. Además, en ambos ejemplos se proporciona el o los artículos de donde fue extraída la información con la que se generó la respuesta.

Sin embargo, en 2 de los 15 casos de evaluación el modelo no logra generar respuestas del todo precisas. Por ejemplo, la respuesta a la pregunta: ¿Qué es un delito doloso?”, cuya respuesta no puede ser encontrada directamente en la constitución, es la siguiente:

*“Un delito doloso es aquel en el que el autor actúa con intención de cometer el delito, es decir, tiene la voluntad de realizar la conducta delictiva y de causar el resultado que se tipifica como delito en la ley.*

*Fuente: Artículo 14 de la Constitución de México.”*

Aunque la definición de delito doloso generada por el modelo es correcta, el artículo 14 de nuestra constitución no incluye esta definición entre sus apartados. Sin embargo, dado que este artículo tiene el contexto más parecido al de la pregunta, el modelo cita su contenido al generar la respuesta.

Por este tipo de inconvenientes sugerimos como trabajo futuro implementar técnicas para la detección de consultas sin el contexto adecuado para generar una respuesta, esto con el fin de evitar la generación espontánea de respuestas no fundamentadas. De igual manera se sugiere la integración de los diversos reglamentos y códigos que complementan nuestra constitución, así como las leyes estatales para obtener la respuesta de la fuente adecuada.

El método implementado se encuentra disponible en <https://github.com/daniel-lima-lopez/Mex-IA-a-Retrieval-Augmented-Generation-Chatbot>, donde se incluye el resto de preguntas de prueba ingresadas al chatbot.

## Referencias

- [1] Anthony Kay. Tesseract: an open-source optical character recognition engine. *Linux Journal*, 2007(159):2, 2007.
- [2] Jerry Liu. Llamaindex. [https://github.com/jerryjliu/llama\\_index](https://github.com/jerryjliu/llama_index), 2022.
- [3] World Justice Project. La mitad de los mexicanos tiene problemas legales, solo uno de cada cuatro tiene ayuda, 2024. Acceso a la justicia en México: desafíos y estadísticas sobre problemas legales en la población.