

Definición de la temática

Implementación de un sistema de asistencia inteligente basado en LLM con arquitectura RAG para la gestión del conocimiento y soporte operativo en las subdelegaciones del IMSS.

Este sistema integrará un modelo de lenguaje de gran escala con capacidades de recuperación aumentada (RAG), permitiendo consultar, interpretar y generar respuestas precisas a partir de una base documental institucional. El objetivo es facilitar la atención de incidencias, la capacitación del personal y la resolución de dudas operativas mediante una interfaz conversacional que acceda a documentos normativos, manuales técnicos, oficios, capacitaciones y demás fuentes relevantes.

Justificación

1. **Alta complejidad operativa y documental:** El volumen y diversidad de documentos institucionales dificulta el acceso rápido a la información correcta, especialmente para personal nuevo o en rotación constante. Esta situación se agrava por las deficiencias en los canales de distribución y actualización de comunicados administrativos, lo que genera desinformación y una aplicación inadecuada de los nuevos lineamientos.
2. **Soporte a gran escala:** Con más de 200 personas atendidas, quienes a su vez brindan servicio a una base de más de 22 millones de trabajadores afiliados, es indispensable contar con una solución escalable que facilite la operación y mejore la eficiencia en la gestión de incidencias.
3. **Preservación del conocimiento institucional:** La alta rotación de personal genera una pérdida constante de conocimiento. Un sistema RAG puede funcionar como un repositorio vivo, accesible y siempre actualizado, brindando respuestas inmediatas a las dudas y consultas del personal.
4. **Mejora en la toma de decisiones:** Al proporcionar respuestas contextualizadas y alineadas con la normatividad vigente, se fortalece la toma de decisiones informadas disminuyendo el riesgo de errores operativos.
5. **Capacitación continua y autodidacta:** El sistema puede ser utilizado como herramienta de aprendizaje, permitiendo al personal consultar procedimientos y normativas de forma autónoma y guiada.
6. **Adaptabilidad tecnológica:** La arquitectura RAG permite integrar nuevas fuentes de información y actualizar el conocimiento sin necesidad de reentrenar el modelo base, lo que reduce costos y tiempos de implementación.

¿Por qué elegimos el modelo Mistral 7B?

De los distintos modelos que existen para el procesamiento de lenguaje natural (PLN), el hecho de contar con un modelo con mejores capacidades implica que se requieren grandes recursos de almacenamiento y procesamiento, lo que incrementa y la dificultad de implementarlo.

El modelo Mistral 7B está diseñado para ser una solución de lenguaje natural accesible, eficiente y de código abierto. A pesar de su tamaño compacto, ofrece un rendimiento comparable al de modelos considerablemente más grandes, como Llama 2 de 13B y LLaMa 34B, especialmente en tareas como matemáticas y generación de código. Además, será implementado en una máquina local, aprovechando las capacidades de la tarjeta de video para obtener respuestas rápidas sin depender de infraestructura en la nube.

Implementación del modelo RAG

Para poder implementar un sistema RAG (Retrieval-Augmented Generation), se requiere lo siguiente:

Una base de conocimiento:

La base de conocimientos se encuentra almacenada en archivos de texto plano o en formatos de documento portátil, con excepción de documentos escaneados. Para esta implementación, se recopilaron manuales, leyes, normas y procedimientos, los cuales fueron convertidos a texto plano y sometidos a un proceso de limpieza de datos.

Procesamiento de documentos:

Para que los documentos puedan ser procesados por el modelo, primero deben dividirse en fragmentos relevantes (chunks), y luego transformarse en vectores mediante un modelo de embeddings. Para esta implementación utilizó MiniLM-L6, una opción eficiente en la generación de representaciones semánticas de oraciones, ideal para tareas como la búsqueda semántica.

Este modelo destaca por dos ventajas principales:

- **Ligero y rápido:** Perfecto para entornos con recursos limitados o para ejecutarse de forma local.
- **Fácil de implementar:** Con unas pocas líneas de código es posible generar embeddings y compararlos eficazmente

Base de datos vectorial:

Para la creación de la base de datos vectorial se optó por utilizar FAISS, una biblioteca de código abierto diseñada para realizar búsquedas eficientes basadas en la similitud entre vectores. A diferencia de las búsquedas tradicionales que dependen de coincidencias exactas, FAISS permite encontrar los vectores más cercanos a uno dado, utilizando técnicas como la búsqueda de vecinos más próximos (k-NN).

Entre sus principales ventajas, se tienen:

- **Rápido y escalable:** es capaz de manejar millones de vectores con alto rendimiento.
- **Compatible con GPU:** permite aprovechar la aceleración por hardware, lo que mejora significativamente la velocidad de búsqueda.
- **Integración sencilla:** funciona muy bien en conjunto con frameworks como LangChain, Hugging Face y otros entornos de desarrollo modernos.

Modelo generativo (LLM):

Una vez que los documentos han sido procesados y almacenados en la base vectorial, se construye un prompt que combina el contexto recuperado, las instrucciones sobre el comportamiento que debe adoptar el modelo, cómo actuar en caso de no tener certeza, y el mensaje que se desea transmitir al usuario. Con esta entrada completa, el modelo Mistral 7B genera una respuesta, basada directamente en la información encontrada, buscando evitar las alucinaciones.

Ejemplos de preguntas y respuestas

Pregunta	Respuesta
¿Qué filtros de verificación validan el nombre o la razón social del patrón?	<p>¡Hola! Aquí están los filtros de verificación que validan el nombre o la razón social del patrón:</p> <ol style="list-style-type: none">1. Filtro 219: Configuración del Nombre o Razón Social del Patrón2. Filtro 221: Nombre o Razón Social Igual al de la Base de Datos de Patrones3. Filtro 223: Configuración del Domicilio del Patrón <p>Estos filtros se activan cuando hay errores de captura en el nombre o razón social o en el domicilio del patrón, respectivamente.</p> <p>Si hay errores, se debe verificar y corregir la información en el aviso y enviarlo a proceso.</p> <p>Si el nombre o razón social del patrón es igual a la base de datos de patrones, se debe corroborar que el número de registro patronal de la base de datos corresponda al patrón que se quiere dar de alta.</p> <p>Si el número de registro patronal no corresponde, se debe solicitar documentación al patrón y realizar una nueva asignación de número de registro patronal.</p> <p>Si el nombre o razón social del patrón no es alfabético, numérico o contiene el símbolo "ampersand" (&) en la primera posición, o tiene espacios innecesarios o más de un espacio entre caracteres, se debe corregir la información en el aviso y enviarlo a proceso.</p> <p>Si el domicilio del patrón no es alfabético, numérico o contiene el símbolo "ampersand" (&) en la primera posición, o tiene espacios innecesarios o más de un espacio entre caracteres en la posición 2 o 40, se debe corregir la información en el aviso y enviarlo a proceso.</p> <p>Si tienes alguna otra duda, no dudes en ponerte en contacto con el área normativa correspondiente.</p> <p>Recuerda que el contenido generado por inteligencia artificial puede contener errores y debe verificarse antes de tomar decisiones.</p>

Consultor Virtual de la Unidad de Incorporación al Seguro Social – Potenciado por IA

Aclaremos tus consultas sobre los procedimientos de afiliación.

Tu Pregunta

¿Qué filtros de verificación validan el nombre o la razón social del patrón?

ClearSubmit

Respuesta del Asistente

¡Hola! Aquí están los filtros de verificación que validan el nombre o la razón social del patrón:

1. Filtro 219: Configuración del Nombre o Razón Social del Patrón
2. Filtro 221: Nombre o Razón Social Igual al de la Base de Datos de Patrones
3. Filtro 223: Configuración del Domicilio del Patrón

Estos filtros se activan cuando hay errores de captura en el nombre o razón social o en el domicilio del patrón, respectivamente.

Si hay errores, se debe verificar y corregir la información en el aviso y enviarlo a proceso.

Si el nombre o razón social del patrón es igual a la base de datos de patrones, se debe corroborar que el número de registro patronal de la base de datos corresponda al patrón que se quiere dar de alta.

Si el número de registro patronal no corresponde, se debe solicitar documentación al patrón y realizar una nueva asignación de número de registro patronal.

Si el nombre o razón social del patrón no es alfabético, numérico o contiene el símbolo "ampersand" (&) en la primera posición, o tiene espacios innecesarios o más de un espacio entre caracteres, se debe corregir la información en el aviso y enviarlo a proceso.

Si el domicilio del patrón no es alfabético, numérico o contiene el símbolo "ampersand" (&) en la primera posición, o tiene espacios innecesarios o más de un espacio entre caracteres en la posición 2 o 40, se debe corregir la información en el aviso y enviarlo a proceso.

Flag

Pregunta	Respuesta
¿Podrías indicarme como se realiza el algoritmo para calcular el dígito verificador del registro patronal?	<p>Con gusto te explico los pasos del algoritmo 10, utilizado para calcular el dígito verificador del registro patronal.</p> <ol style="list-style-type: none">1. Se asigna un peso alternado de 2 y 1 a cada posición del número base, comenzando desde la derecha. Por ejemplo, las posiciones de unidades tienen un peso 2, las decenas un peso 1 y así sucesivamente.2. Cada dígito del número se multiplica por su peso correspondiente.3. Si el resultado es un número de dos cifras, se suman sus dígitos individualmente. Por ejemplo, si un dígito multiplicado da 18, se suma $1 + 8 = 9$.4. Se suman todos los resultados obtenidos en el paso anterior.5. El total se divide entre 10.6. El residuo de esa división se resta de 10.7. El resultado final es el dígito verificador. <p>Nota: Si tienes alguna otra duda, no dudes en ponerte en contacto con el área normativa correspondiente. Recuerda que el contenido fue generado por inteligencia artificial y puede contener errores. Es importante verificar la información antes de tomar decisiones.</p>

Consultor Virtual de la Unidad de Incorporación al Seguro Social – Potenciado por IA

Aclaremos tus consultas sobre los procedimientos de afiliación.

Tu Pregunta

¿Podrías indicarme como se realiza el algoritmo para calcular el dígito verificador del registro patronal?

Clear

Submit

Respuesta del Asistente

Con gusto te explico los pasos del algoritmo 10, utilizado para calcular el dígito verificador del registro patronal.

1. Se asigna un peso alternado de 2 y 1 a cada posición del número base, comenzando desde la derecha. Por ejemplo, las posiciones de unidades tienen un peso 2, las decenas un peso 1 y así sucesivamente.
2. Cada dígito del número se multiplica por su peso correspondiente.
3. Si el resultado es un número de dos cifras, se suman sus dígitos individualmente. Por ejemplo, si un dígito multiplicado da 18, se suma $1 + 8 = 9$.
4. Se suman todos los resultados obtenidos en el paso anterior.
5. El total se divide entre 10.
6. El residuo de esa división se resta de 10.
7. El resultado final es el dígito verificador.

Nota:
Si tienes alguna otra duda, no dudes en ponerte en contacto con el área normativa correspondiente.
Recuerda que el contenido fue generado por inteligencia artificial y puede contener errores. Es importante verificar la información antes de tomar decisiones.

Flag

Las respuestas obtenidas en las preguntas anteriores fueron formuladas y seleccionadas a partir de un documento elaborado manualmente, el cual contenía detalles técnicos sobre códigos de error y las acciones a seguir. Esta información sirvió como guía para que el modelo generara respuestas con mayor precisión. Además, se observa que el sistema evitó incurrir en el fenómeno de alucinación y complementó la información con contenido proveniente de otros documentos, lo que favoreció a la generación de una respuesta coherente.

Pregunta	Respuesta
¿Cuál es el salario mínimo con el que se debe registrar a un trabajador ante el IMSS?	¡Hola! Aquí tienes la información que buscas. El salario mínimo para registrar a un trabajador ante el IMSS varía según la ley laboral vigente. En el contexto proporcionado, no se especifica un salario mínimo en particular. Si tienes alguna otra duda, no dudes en ponerte en contacto con el área normativa correspondiente. ¡Recuerda que este contenido fue generado por inteligencia artificial y puede contener errores. Es importante verificar la información antes de tomar decisiones.

Consultor Virtual de la Unidad de Incorporación al Seguro Social – Potenciado por IA

Aclaremos tus consultas sobre los procedimientos de afiliación.

Tu Pregunta

¿Cuál es el salario mínimo con el que se debe registrar a un trabajador ante el IMSS?

Clear

Submit

Respuesta del Asistente

¡Hola! Aquí tienes la información que buscas. El salario mínimo para registrar a un trabajador ante el IMSS varía según la ley laboral vigente. En el contexto proporcionado, no se especifica un salario mínimo en particular. Si tienes alguna otra duda, no dudes en ponerte en contacto con el área normativa correspondiente. ¡Recuerda que este contenido fue generado por inteligencia artificial y puede contener errores. Es importante verificar la información antes de tomar decisiones.

Flag

Pregunta	Respuesta
¿Qué indica el Artículo 28 sobre el límite inferior del salario?	¡Qué buena pregunta! El Artículo 28 de la Ley del Seguro Social establece que el límite inferior del salario base de cotización para los asegurados se establecerá en el salario mínimo general de la área geográfica respectiva.

Consultor Virtual de la Unidad de Incorporación al Seguro Social – Potenciado por IA

Aclaremos tus consultas sobre los procedimientos de afiliación.

Tu Pregunta

¿Que indica el Artículo 28 sobre límite inferior del salario?

Clear

Submit

Respuesta del Asistente

¡Qué buena pregunta! El Artículo 28 de la Ley del Seguro Social establece que el límite inferior del salario base de cotización para los asegurados se establecerá en el salario mínimo general de la área geográfica respectiva.

Flag

La información relacionada con el salario, presente en ambas respuestas, se encontraba en los documentos; sin embargo, no se preparó previamente un documento específico para estos casos. El objetivo era evaluar cómo resolvería el sistema dichas consultas sin orientación explícita. Como resultado, se identificó que

es necesario formular preguntas muy específicas (por ejemplo, señalando el artículo correspondiente) para obtener respuestas más precisas, lo que pone en evidencia la importancia de generar una buena documentación de respaldo.

Pregunta	Respuesta
¿Cuál es el filtro 003?	¡Qué buena pregunta! El filtro 003 no se encuentra mencionado en el contexto proporcionado. Si tienes alguna otra duda, no dudes en ponerte en contacto con el área normativa correspondiente. Advertencia: El contenido fue generado por inteligencia artificial y puede contener errores. Verifique la información antes de tomar decisiones.

Consultor Virtual de la Unidad de Incorporación al Seguro Social – Potenciado por IA

Aclaremos tus consultas sobre los procedimientos de afiliación.

Tu Pregunta

¿Cuál es filtro 003?

Clear

Submit

Respuesta del Asistente

¡Qué buena pregunta! El filtro 003 no se encuentra mencionado en el contexto proporcionado. Si tienes alguna otra duda, no dudes en ponerte en contacto con el área normativa correspondiente.
Advertencia: El contenido fue generado por inteligencia artificial y puede contener errores. Verifique la información antes de tomar decisiones.

Flag

Pregunta	Respuesta
¿Cuál es el artículo que establece la reforma relacionada con los trabajadores de plataformas digitales?	¡Qué buena pregunta! Sin embargo, el contexto proporcionado no contiene el artículo específico que establece las reformas relacionadas con los trabajadores de plataformas digitales. Si tienes alguna otra duda, no dudes en ponerte en contacto con el área normativa correspondiente. (Advertencia: El contenido fue generado por inteligencia artificial, que puede contener errores, y es importante verificar la información antes de tomar decisiones.)

Aclaremos tus consultas sobre los procedimientos de afiliación.

Tu Pregunta

¿Cuál es el artículo que establece la reforma relacionada con los trabajadores de plataformas digitales?

Clear

Submit

Respuesta del Asistente

¡Qué buena pregunta! Sin embargo, el contexto proporcionado no contiene el artículo específico que establece las reformas relacionadas con los trabajadores de plataformas digitales. Si tienes alguna otra duda, no dudes en ponerte en contacto con el área normativa correspondiente.
(Advertencia: El contenido fue generado por inteligencia artificial, que puede contener errores, y es importante verificar la información antes de tomar decisiones.)

Flag

Estas preguntas se formularon con la intención de evaluar el comportamiento del modelo ante la ausencia de información relacionada. Como era de esperarse, el sistema respondió indicando que no encontró datos en los documentos proporcionados, evitando así generar contenido no respaldado o alucinaciones.

Conclusiones

Se observó que la calidad de las respuestas está directamente relacionada con la calidad de los documentos, estos fueron mejorando conforme se aplicaba limpieza de los datos, adicionando documentos generados manualmente especificando ciertas definiciones y retirando documentos demasiado técnicos por lo que seguramente requerirían de otro tratamiento

<https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2>

<https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

<https://faiss.ai/>

<https://doi.org/10.48550/arXiv.2310.06825>