DESDE 2013
https://repository.uaeh.edu.mx/revistas/index.php/icbi/issue/archive
Pädi Boletín Científico de Ciencias Básicas e Ingenierías del ICBI

ISSN: 2007-6363

Publicación Semestral Pädi Vol. 13, No. Especial 2 (2025) 39 - 47

Evaluación de coherencia en respuestas generadas por un Chatbot para consulta de legislación y reglamentos universitarios
Evaluation of coherence in chatbot-generated responses for querying university legislation and regulations

C. M. Santibáñez-Camarillo [a], C. E. Millán-Hernández [b], E. Sánchez-Soto [c]

[a] *División de Estudios de Posgrado, Universidad Tecnológica de la Mixteca, 69004, Huajuapan de León, Oaxaca, México.*
[b,c] *Instituto de Computación, Universidad Tecnológica de la Mixteca, 69004, Huajuapan de León, Oaxaca, México.*

Resumen

Los organismos públicos suelen contar con numerosos reglamentos que abarcan diversos ámbitos administrativos y operacionales. Este volumen de documentos puede dificultar que los interesados estén al tanto de todas las normativas, así como la accesibilidad a la información relevante. Para abordar este desafío, se propone un sistema basado en un agente de preguntas y respuestas, para que los usuarios realicen consultas a los reglamentos institucionales. Este sistema utiliza Generación Aumentada por Recuperación (RAG), identificando fragmentos relevantes en los reglamentos, generando respuestas informativas y contextualizadas en tiempo real. En este artículo se propone la evaluación de las respuestas mediante la combinación de las métricas *BLEU* y *ROUGE*, para medir la similitud con respuestas de referencia, en adición con *Perplexity* para cuantificar la coherencia del texto. Los resultados de las consultas realizadas a través del sistema propuesto facilitan el acceso a la normativa y optimizan el tiempo de consulta, propiciando un entorno de trabajo más eficiente y con mejor acceso a la información.

Palabras clave: Generación Aumentada por Recuperación, Procesamiento de Lenguaje Natural, BLEU, ROUGE, Perplejidad, Sistema de preguntas y respuestas.

Abstract

Public organizations often have numerous regulations that cover various administrative and operational areas. This volume of documents can make it difficult for stakeholders to be aware of all the regulations, as well as to access relevant information. To address this challenge, a system based on a question-and-answer agent is proposed, allowing users to query institutional regulations. This system uses Retrieval-Augmented Generation (RAG), identifying relevant fragments in the regulations and generating informative and contextualized responses in real time. The accuracy of the responses is evaluated using BLEU and ROUGE metrics to measure similarity to reference answers. Additionally, Perplexity is used to quantify the coherence of the text. The results of the queries made through the proposed system facilitate access to the regulations and optimize query time, fostering a more efficient work environment with better access to information.

Keywords: Retrieval-Augmented Generation, Natural Language Processing, BLEU, ROUGE, Perplexity, Question-and-answer system.

1. Introducción

En los ámbitos académico y administrativo, el acceso eficiente a la información es fundamental para la toma de decisiones fundamentadas y para promover un entorno operativo eficaz. Sin embargo, muchas instituciones enfrentan el desafío de gestionar un volumen considerable de normativas y reglamentos que regulan sus actividades. Este hecho dificulta que los usuarios accedan de manera rápida y precisa a la normativa relevante, lo que afecta la eficiencia en los procesos de consulta y cumplimiento.

El propósito de este proyecto es diseñar e implementar un agente de preguntas y respuestas (QAa, por las siglas en inglés *Question Answering agent*), permitiendo a los usuarios consultar los reglamentos institucionales y obtener información clara, precisa y contextualizada en tiempo real. Para alcanzar este objetivo, se propone analizar y clasificar los reglamentos para identificar categorías clave, construir un

*Autor para la correspondencia: carlosmsanti@gs.utm.mx
Correo electrónico: carlosmsanti@gs.utm.mx (Carlos Manuel Santibáñez-Camarillo), ceduardo.millan@gmail.com (Christian Eduardo Millán-Hernández), esanchez@mixteco.utm.mx (Eduardo Sánchez-Soto).

Historial del manuscrito: recibido el 16/03/2025, última versión-revisada recibida el 27/06/2025, aceptado el 31/07/2025, publicado el 15/09/2025. **DOI:** https://doi.org/10.29057/icbi.v13iEspecial2.14676

corpus documental que facilite la búsqueda eficiente de información relevante e implementar un sistema de Generación Aumentada por Recuperación RAG (Por sus siglas en inglés *Retrieval Augmented Generation*) que funcione como un corpus adaptado específicamente para responder preguntas relacionadas con las normativas institucionales. Además, la calidad de las respuestas generadas por el sistema a los cuestionamientos realizados por un usuario sobre los reglamentos de algún organismo público, será evaluada mediante métricas estándar de Procesamiento del Lenguaje Natural (PLN) como BLEU, ROUGE y Perplexity, para asegurar la precisión y coherencia lingüística.

Este trabajo se fundamenta en teorías y técnicas avanzadas de PLN, integrando métodos de recuperación de información y generación automática de texto para garantizar la relevancia y utilidad de las respuestas. Asimismo, se apoya en un marco conceptual basado en modelos de lenguaje preentrenados, como *LlamaIndex*, una herramienta que facilita la indexación y recuperación de información desde fuentes externas para personalizar respuestas en aplicaciones específicas (LlamaIndex, 2024). Este enfoque permite optimizar el acceso a normativas y garantizar la relevancia de las respuestas generadas por el sistema, lo que facilita la personalización de las respuestas y asegura que los datos utilizados sean confiables y contextualizados para aplicaciones específicas. Por ejemplo, los *chatbots* en administraciones públicas, como los desarrollados en España, ofrecen acceso a normativas y políticas relevantes, mejorando la eficiencia en la gestión de datos abiertos (Administración Electrónica, 2022). En Argentina, Boti, el *chatbot* de la Ciudad de Buenos Aires, facilita consultas sobre normativas de salud y educación, proporcionando información rápida y contextualizada (Gobierno de la Ciudad de Buenos Aires, 2024). Además, los QAa extraen información precisa de documentos extensos, siendo útiles en contextos normativos y académicos (Laparra, 2024). Copilot.Live personaliza soluciones para gobiernos, optimizando el acceso a normativas con chatbots avanzados (Copilot.Live, 2024). Por su parte, TEO, el chatbot de la Fiscalía de Oaxaca, facilita trámites legales, combate la corrupción y permite registrar quejas y denuncias con una interfaz amigable (El Universal, 2025). Estas herramientas demuestran cómo la tecnología puede adaptarse a necesidades específicas, garantizando eficiencia y confiabilidad en la gestión de información. Sin embargo, no se cuenta con una evaluación de la calidad de las respuestas generadas en los estudios antes mencionados (Administración Electrónica, 2022; Copilot.Live, 2024; El Universal, 2025; Gobierno de la Ciudad de Buenos Aires, 2024).

2. **Marco Teórico**

El marco conceptual de esta investigación se centra en la implementación de QAa integrado con modelos de lenguaje preentrenados, diseñados para optimizar el acceso a normativas en organismos públicos. Los QAa emplean arquitecturas de aprendizaje profundo, como *transformers*, que han demostrado un desempeño sobresaliente en tareas de Procesamiento del Lenguaje Natural (PLN). Ejemplos destacados incluyen *LlamaIndex*, *Cohere*, GPT y *LangChain*, cada uno con capacidades específicas que enriquecen la funcionalidad de estos sistemas.

Cohere se especializa en la creación de modelos de lenguaje orientados a equipos y empresas, ofreciendo servicios para clasificación, generación de texto y análisis de sentimiento en múltiples idiomas. Su enfoque permite personalizar modelos de lenguaje adaptados a necesidades específicas, mejorando la precisión de las respuestas (Cohere, 2025).

LangChain, por otro lado, proporciona un marco de desarrollo que conecta modelos de lenguaje con fuentes de datos externas, como bases de datos y APIs, permitiendo la generación de respuestas contextuales y precisas en entornos complejos. Su versatilidad lo hace ideal para construir aplicaciones de PLN integradas con flujos de datos personalizados (LangChain, 2025).

OpenAI, a través de su modelo GPT-4, combina capacidades avanzadas de generación de texto con técnicas de razonamiento multietapa, lo que resulta en un desempeño sobresaliente en tareas como la respuesta a preguntas complejas y el análisis de información contextual. Este modelo es ampliamente utilizado en aplicaciones de chatbots y generación de contenido especializado (OpenAI, 2025).

Además de las opciones anteriores, *DeepSeek*, es una plataforma emergente que integra técnicas avanzadas de PLN y RAG para mejorar la precisión y relevancia de las respuestas generadas. Su enfoque se centra en la optimización de la recuperación de información y la generación de respuestas contextualizadas, lo que lo convierte en una herramienta valiosa para aplicaciones en organismos públicos y entornos normativos (DeepSeek, 2024).

En conjunto, estas tecnologías representan un avance significativo en la capacidad de los sistemas para procesar y generar información relevante, garantizando un acceso eficiente y confiable a normativas y datos en contextos específicos.

Un ejemplo destacado de este marco de trabajo es *LlamaIndex*, que actúa como interfaz entre fuentes de datos estructuradas (como reglamentos) y modelos de lenguaje. Este marco de trabajo permite:

1. Extracción de Información Relevante: El sistema identifica y extrae fragmentos específicos de texto relacionados con las consultas de los usuarios mediante técnicas de recuperación de información.
2. Generación de Respuestas Contextualizadas: Una vez identificada la información relevante, el modelo genera respuestas claras, coherentes y adaptadas al contexto.
3. Personalización y Escalabilidad: Los modelos pueden ajustarse a los contextos específicos de los organismos públicos, asegurando respuestas más relevantes y pertinentes. Además, la escalabilidad permite que estos modelos manejen grandes volúmenes de información y consultas simultáneas sin degradar su rendimiento, facilitando su implementación en distintas dependencias o niveles de gobierno.

LlamaIndex es clave en este marco, ya que facilita la indexación de datos y la generación de *embeddings* para la recuperación eficiente de información. Además, minimiza las limitaciones tradicionales de los modelos de lenguaje preentrenados al integrar datos externos confiables, reduciendo la probabilidad de generar respuestas irrelevantes o incorrectas.

Diversas implementaciones en contextos públicos respaldan la viabilidad de los sistemas RAG y *chatbots* como

herramientas para democratizar el acceso a información normativa:

- *Chatbots* en administraciones públicas: Herramientas implementadas para facilitar el acceso a datos abiertos, como los desarrollados en España para ofrecer normativas y políticas relevantes (Administración Electrónica, 2022).
- Boti, el chatbot de la Ciudad de Buenos Aires: Este chatbot facilita la consulta de normativas sobre salud y educación, proporcionando información rápida y contextualizada (Gobierno de la Ciudad de Buenos Aires, 2024).
- Sistemas RAG para extracción de información: Estas técnicas se emplean ampliamente para generar respuestas claras y precisas a partir de documentos extensos en contextos normativos y académicos (Laparra, 2024).
- Copilot.Live: Una plataforma especializada en la personalización de soluciones para gobiernos, optimizando el acceso a normativas mediante chatbots avanzados (Copilot.Live, 2024).
- TEO, el chatbot de la Fiscalía General del Estado de Oaxaca: Este sistema, diseñado para mejorar el acceso a trámites legales y combatir la corrupción, destaca por su interfaz amigable y su capacidad para responder preguntas frecuentes, además de registrar quejas y denuncias (El Universal, 2025).

Estos ejemplos reflejan cómo las tecnologías basadas en RAG y *chatbots* están transformando la interacción entre ciudadanos y normativas, optimizando la accesibilidad y eficiencia en diversos contextos públicos.

El sistema propuesto en esta investigación busca replicar y adaptar estas estrategias al ámbito de los organismos públicos, empleando técnicas avanzadas de PLN para transformar la forma en que se accede a los reglamentos. Este enfoque no solo facilita la búsqueda de información relevante, sino que también genera respuestas claras y contextualizadas, fortaleciendo la capacidad de decisión y el cumplimiento normativo dentro de las instituciones.

Según Greyling (2023), los agentes basados en modelos de lenguaje de gran escala (*LLM*) han demostrado su capacidad para realizar tareas de respuesta a preguntas complejas mediante razonamiento multinivel, lo cual mejora la precisión y relevancia de las respuestas generadas.

2.1. *Evolución del Procesamiento del Lenguaje Natural (PLN)*

El campo del PLN ha evolucionado significativamente con la incorporación de arquitecturas de aprendizaje profundo como los *transformers* (Vaswani et al., 2017). Estas arquitecturas son la base de los modelos de lenguaje actuales, como GPT-4, que combinan recuperación de información y generación de texto. Según Manning y Schütze (1999, citado por Takahashi y Tanaka-Ishii, 2019), métricas como Perplejidad, BLEU y ROUGE son fundamentales para medir la calidad de estos sistemas, especialmente en contextos académicos y normativos.

Además, el uso de *embeddings* vectoriales (por ejemplo, mediante *FAISS*, desarrollado por *Facebook AI*) permite realizar búsquedas más eficientes en bases de datos de gran escala. Esto es clave para indexar normativas institucionales y responder preguntas de manera precisa en tiempo real (Laparra, 2024).

Desde un punto de vista teórico, este proyecto se fundamenta en conceptos clave del PLN y en la evaluación de modelos de lenguaje mediante métricas como BLEU, ROUGE y Perplexity, (Perplejidad) garantizando la calidad y coherencia de las respuestas generadas. Con esta investigación, se busca abordar un problema práctico y contribuir al desarrollo de herramientas tecnológicas que optimicen el acceso a la información en organismos públicos, promoviendo un entorno operativo más eficiente y mejor informado.

Los modelos de lenguaje se evalúan tradicionalmente mediante dos enfoques principales: el análisis directo de sus componentes internos o la verificación de los resultados generados por el modelo (Manning & Schütze, 1999, citado por Takahashi & Tanaka-Ishii, 2019). El primer enfoque se basa en métricas como la distribución de probabilidad, siendo la perplejidad una métrica estándar en modelos *n-gramas* y redes neuronales. Esta mide la precisión en la predicción de palabras considerando el contexto, proporcionando una evaluación cuantitativa del rendimiento de un modelo. Por ejemplo, la perplejidad se calcula en (1).

$$\text{Perplejidad} = e^{-\frac{1}{N}\sum_{i=1}^{N}\log q(x_i)} \quad (1)$$

Donde $q(x_i)$ representa la probabilidad predicha para cada palabra (Manning & Schütze, 1999, citado por Takahashi & Tanaka-Ishii, 2019). La perplejidad es una métrica que mide la coherencia y fluidez del texto generado. Según Patel (2023), la perplejidad es una métrica clave para evaluar la calidad de los modelos de lenguaje, ya que cuantifica qué tan bien el modelo predice la siguiente palabra en una secuencia. Cuanto menor sea la perplejidad, más probable es que el texto haya sido generado por un modelo de lenguaje bien entrenado. Se calcula como la exponencial de la pérdida (loss) del modelo de lenguaje. Un valor bajo indica que el modelo predice correctamente las siguientes palabras en una secuencia, lo que sugiere que el texto generado es coherente y fluido. En este proyecto, la perplejidad se utiliza para evaluar la calidad lingüística de las respuestas generadas, asegurando que sean coherentes y naturales.

El segundo enfoque utiliza referencias para evaluar resultados generados. BLEU (Bilingual Evaluation Understudy) es una métrica que evalúa la similitud entre el texto generado y una referencia (respuesta esperada) mediante el análisis de n-gramas (secuencias de palabras). Métricas como BLEU (Papineni et al., 2002, citado por Takahashi & Tanaka-Ishii, 2019) y ROUGE (Lin, 2004, citado por Takahashi & Tanaka-Ishii, 2019) comparan los n-gramas generados por el modelo con respuestas expertas. Aunque son comunes en tareas como la traducción y el resumen automático, su utilidad es limitada a casos donde existen pares de texto y referencia, además de mostrar inconsistencias en algunos escenarios (Takahashi & Tanaka-Ishii, 2019).

Según Patel (2023), BLEU es especialmente útil para medir la precisión en la traducción automática y la generación de texto, ya que compara la coincidencia de n-gramas entre el texto generado y la referencia. BLEU compara la precisión de los n-gramas (palabras individuales, pares de palabras, etc.) en el texto generado con los de la referencia. Se aplica un factor de penalización para textos más cortos, ya que pueden tener una coincidencia accidentalmente alta. En este proyecto, BLEU se utiliza para medir cuán cercana es la respuesta generada a la respuesta esperada en términos de estructura y elección de palabras.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) es una métrica orientada a la recuperación de información, que mide la coincidencia entre el texto generado y la referencia en términos de palabras clave y secuencias de palabras. Patel (2023) destaca que ROUGE es particularmente útil en tareas de resumen automático, donde la recuperación de información relevante es crucial. ROUGE se calcula en varias versiones:

ROUGE-1: Coincidencias de unigramas (palabras individuales).

ROUGE-2: Coincidencias de bigramas (pares de palabras).

ROUGE-L: Longitud de la subsecuencia más larga común (LCS, por sus siglas en inglés), que mide la coincidencia de la secuencia más larga de palabras que aparece tanto en el texto generado como en la referencia.

ROUGE se utiliza para evaluar la precisión en la recuperación de palabras clave y la estructura de las respuestas generadas.

Por otro lado, métodos alternativos incluyen la evaluación mediante otros modelos de lenguaje. Por ejemplo, Fedus et al. (2018, citado por Takahashi & Tanaka-Ishii, 2019) proponen analizar texto generado por GANs mediante modelos neuronales entrenados en el mismo conjunto de datos. Aunque prometedora, esta aproximación depende de la confiabilidad del modelo evaluador y enfrenta limitaciones en la representación de estructuras lingüísticas complejas. Adicionalmente, se han explorado métodos como el uso de gramáticas probabilísticas (PCFG), las cuales, aunque útiles para análisis estructurales, tienen limitaciones para evaluar aspectos gramaticales más amplios (Takahashi & Tanaka-Ishii, 2019).

Esta diversidad de métricas refleja la complejidad inherente a la evaluación de modelos de lenguaje, destacando la importancia de seleccionar métodos que correspondan al propósito específico del modelo analizado.

Estas métricas son fundamentales para evaluar la calidad de los sistemas de generación de lenguaje, ya que permiten cuantificar tanto la precisión semántica como la coherencia del texto generado. La combinación de BLEU, ROUGE y Perplexity proporciona una evaluación integral del desempeño del sistema, asegurando que las respuestas no solo sean precisas en términos de contenido, sino también coherentes y fluidas en términos lingüísticos.

3. **Caso de Estudio**

La Universidad Tecnológica de la Mixteca (UTM) dispone de una extensa colección de normativas que regulan su funcionamiento, organizadas de manera estructurada para abarcar distintas áreas de la vida académica y administrativa. Estas normativas están compuestas por diferentes niveles de detalle:

- Capítulos: Divisiones principales que agrupan los reglamentos según áreas temáticas o funcionales.
- Artículos: Secciones específicas que detallan normativas, derechos y responsabilidades.
- Secciones y subsecciones: Subdivisiones dentro de los artículos que desarrollan puntos concretos de la normativa.
- Apéndices: Material complementario que incluye ejemplos, aclaraciones o formatos relevantes.

3.1. *Análisis del problema y recopilación de reglamentos*

Para abordar el acceso eficiente a esta amplia normativa, se llevó a cabo una etapa inicial que consistió en identificar y recopilar los reglamentos académicos y administrativos de la UTM. Estos documentos fueron clasificados en categorías principales (académicas, administrativas, éticas, operativas y legales), con el objetivo de facilitar su análisis y uso en las siguientes etapas del proyecto.

3.2. *Herramientas utilizadas*

El proceso de recopilación y análisis se apoyó en las siguientes herramientas:

- Software de digitalización y OCR: Para convertir los documentos físicos en un formato digital accesible y procesable.
- Clasificación de documentos: Los documentos recopilados fueron organizados en categorías según su naturaleza y propósito, lo que permitió construir un corpus documental estructurado para la búsqueda y recuperación de información.

La Tabla 1 resume los reglamentos y manuales analizados, agrupados en función de su temática.

Tabla 1. Clasificación de los Reglamentos

Categoría	Ejemplos representativos
Académico	Reglamentos Generales de Posgrado (2010 y 2016), Reglamento de Alumnos (2016 y 2024)
Administrativo	Reglamento de Trabajo, Reglamento de Estímulos a la Carrera Académica, Plan Estratégico Institucional
Ético y de Conducta	Código de Ética, Protocolo contra Hostigamiento y Acoso Sexual, Protocolo de Violencia de Género, Procedimiento de Quejas y Denuncias.
Operativo	Reglamentos de Uso de la Red, Salas y Talleres, Laboratorio de Química, Reglamento de Casas Habitación.
Legal y Normativo	Legislación Universitaria, Decreto de Creación, Resoluciones de la Asamblea General.

El análisis y clasificación de los reglamentos representaron un paso esencial para garantizar que el sistema propuesto pueda acceder y procesar la información de manera eficiente. A partir de este corpus estructurado, se habilitó el desarrollo de un modelo basado en técnicas avanzadas de Procesamiento del Lenguaje Natural, como se detalla en las etapas metodológicas posteriores.

4. **Procedimiento**

El desarrollo del sistema de Recuperación y Generación de Respuestas (RAG) para facilitar el acceso a normativas en la Universidad Tecnológica de la Mixteca (UTM) se realizó siguiendo un conjunto de pasos organizados, representados en el diagrama de flujo mostrado en la Figura 1, que detalla el procedimiento metodológico. Cada uno de los pasos es descrito a continuación:

4.1. *Construcción del corpus documental*

En esta etapa inicial, se utilizó la biblioteca PyPDF2 para extraer el texto de los reglamentos recopilados. Los

documentos fueron convertidos a un formato texto estándar (UTF-8) y su estructura fue normalizada para facilitar el análisis computacional. Se manejaron posibles problemas, como la ausencia de texto en ciertas páginas, garantizando que la extracción fuera consistente y completa.

Figura 1. Procedimiento metodológico

4.2. *División en fragmentos*

Utilizando la herramienta RecursiveCharacterTextSplitter, los textos procesados fueron divididos en fragmentos más pequeños y manejables, optimizando su procesamiento en etapas posteriores.

4.3. *Creación de vectores de incrustación*

Para representar los fragmentos de texto en un formato adecuado para la búsqueda de similitud, se empleó el modelo preentrenado *sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2* de *Hugging Face* (Sentence-Transformers, 2024), que genera representaciones vectoriales del texto. Los vectores generados se almacenaron en una base de datos vectorial utilizando FAISS (Facebook AI Similarity Search), lo que permite realizar búsquedas rápidas y eficientes.

4.4. *Interfaz de usuario*

Se diseñó una interfaz interactiva utilizando *Streamlit* (Streamlit, 2024) que permite a los usuarios cargar documentos, realizar consultas y visualizar las respuestas generadas. Esta interfaz sirve como el punto de acceso principal para interactuar con el sistema.

4.5. *Búsqueda de contexto relevante*

Cuando el usuario realiza una consulta, el sistema utiliza la base de datos vectorial creada en FAISS para buscar los fragmentos más similares al texto de la pregunta. Este proceso garantiza que las respuestas se basen en información relevante y contextual.

4.6. *Generación de respuestas*

Una vez recuperados los fragmentos relevantes, se utiliza la biblioteca *Ollama* para enviar el contexto y la consulta a un modelo de lenguaje avanzado basado en *llama3.2:3b*. Este modelo genera una respuesta precisa y coherente basada en la información recuperada.

En la Figura 1 se resume el flujo metodológico, mostrando cómo cada uno de estos pasos se conecta para construir un sistema eficiente de recuperación y generación de respuestas. Adicionalmente, se desarrolló un índice de búsqueda en FAISS para optimizar las búsquedas de similitud, permitiendo la recuperación ágil de fragmentos relevantes y mejorando la experiencia del usuario.

5. **Implementación de RAG con LlamaIndex**

La elección de LlamaIndex, en combinación con *Ollama*, se fundamenta en tres ventajas principales. En primer lugar, la privacidad es un factor clave, ya que *Ollama* permite ejecutar modelos de lenguaje localmente, garantizando que los datos sensibles permanezcan bajo control total de la organización, sin necesidad de depender de servicios en la nube. En segundo lugar, la eficiencia en la recuperación de información se logra mediante las capacidades avanzadas de indexación y búsqueda de LlamaIndex, que optimizan la localización de fragmentos relevantes dentro de grandes volúmenes de datos normativos. Finalmente, la flexibilidad y personalización permiten adaptar los modelos de lenguaje a contextos específicos, asegurando respuestas precisas y relevantes que consideran las particularidades de cada aplicación.

5.1. *Capacidades de LlamaIndex*

LlamaIndex es un marco diseñado para la implementación de sistemas de RAG en LLM. Este sistema conecta directamente los LLM con fuentes de datos externas, personalizándolos para aplicaciones específicas y reduciendo significativamente la generación de respuestas irrelevantes o incorrectas.

5.2. *Justificación de su selección*

Aunque existen otras herramientas y modelos de lenguaje que podrían emplearse para implementar QAa, *LlamaIndex* fue seleccionado por su capacidad de integración directa con bases de datos y su eficiencia en la recuperación de información. Por ejemplo, *LangChain* es una opción versátil para aplicaciones basadas en LLM, pero carece de la capacidad de indexación avanzada que ofrece *LlamaIndex*. *OpenAI GPT-4*, aunque poderoso, depende de la nube, lo que puede generar problemas de privacidad en ciertos entornos. *Cohere*, por su parte, ofrece soluciones robustas de análisis y generación de texto, pero su falta de integración nativa con bases de datos vectoriales, como *FAISS*, podría limitar su eficiencia en la recuperación de información específica.

Entre las aplicaciones de *LlamaIndex*, destacan los motores de búsqueda personalizados, que facilitan la creación de

sistemas especializados para indexar documentos y realizar consultas específicas. También permite desarrollar *chatbots* especializados que pueden adaptarse a terminología, políticas y conocimientos específicos de un entorno empresarial o institucional. Además, *LlamaIndex* es útil para generar resúmenes precisos y relevantes de documentos extensos, sintetizando información de manera eficiente (Gheorghiu, 2024).

5.3. *Uso de Ollama para la implementación local*

Ollama complementa las capacidades de *LlamaIndex* al permitir ejecutar modelos de lenguaje avanzado de manera local. Esto ofrece un control total sobre los datos, optimiza la gestión de grandes volúmenes de información y elimina la dependencia de la nube. Estas características son especialmente relevantes en entornos donde la privacidad de los datos es prioritaria.

En este proyecto, se empleó el modelo preentrenado *llama3.2:b*, ajustado específicamente para responder preguntas relacionadas con el corpus de reglamentos institucionales. La integración con *FAISS* permitió la recuperación eficiente de fragmentos relevantes y mejoró la generación de respuestas contextualizadas.

6. Desarrollo de la interfaz de usuario

Se desarrolla una interfaz gráfica que permite a los usuarios realizar consultas en lenguaje natural y recibir respuestas contextualizadas. La implementación de esta interfaz se lleva a cabo utilizando la biblioteca *Streamlit*, lo que garantiza una experiencia interactiva y accesible para los usuarios. La interfaz facilita la interacción en tiempo real, asegurando que las consultas sean procesadas de manera eficiente y que las respuestas generadas sean precisas y relevantes.

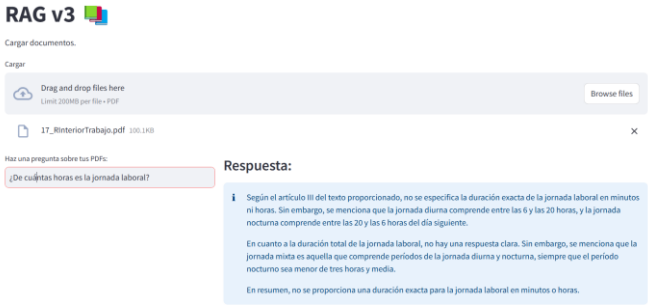

Figura 2. Interfaz de Usuario

En la Figura 2 se muestra la interfaz de usuario desarrollada, que incluye los componentes visuales y las funcionalidades principales del sistema. Para su funcionamiento, se utilizan las siguientes herramientas:

- *Streamlit* para la creación de la interfaz gráfica, que permite una fácil integración con los modelos de lenguaje y proporciona una plataforma interactiva para los usuarios.
- *APIs* de integración que permiten realizar consultas en tiempo real, asegurando que las interacciones con el sistema sean dinámicas y sin interrupciones.

Este desarrollo facilita el acceso a la información contenida en los reglamentos, mejorando la eficiencia en la consulta de normativas y proporcionando respuestas instantáneas y bien contextualizadas.

7. Evaluación del Sistema

Para evaluar la calidad de las respuestas generadas por el sistema de recuperación y generación de texto, se implementó un análisis basado en tres métricas ampliamente utilizadas en el procesamiento de lenguaje natural: BLEU, ROUGE, y Perplejidad. Según Patel (2023), las métricas citadas son esenciales para evaluar la calidad de los modelos de lenguaje, ya que permiten medir tanto la precisión semántica como la coherencia del texto generado. Estas métricas permiten analizar tanto la similitud semántica como la calidad lingüística de las respuestas generadas.

A continuación, se describen los detalles de la implementación.

7.1. *Datos Utilizados*

Se evaluaron 25 preguntas realizadas por estudiantes de la Universidad Tecnológica de la Mixteca relacionadas con normativas universitarias. Estas preguntas fueron seleccionadas a partir de un conjunto inicial de 170 dudas recopiladas en entrevistas del sistema de tutorías, en las que participaron 100 estudiantes de posgrado. La selección se basó en la frecuencia con la que se presentaban las inquietudes, concentrándose en las más recurrentes, lo que resultó en 25 preguntas que representan la totalidad de las dudas recopiladas. A continuación, las preguntas fueron categorizadas con la finalidad de conocer el tipo de normativa descrita en la Tabla 1. Se basó en la recurrencia de las inquietudes presentadas, es decir, se contabilizó la frecuencia agrupando por similitud las preguntas. Adicionalmente, se registró el tiempo de generación de cada respuesta para observar la eficiencia del sistema.

7.2. *Preparación del Modelo y las Métricas*

Se utilizó el modelo preentrenado GPT-2 de la biblioteca Transformers para calcular la Perplejidad, una métrica que mide la coherencia y fluidez del texto generado. Cuanto menor es la perplejidad, mayor es la probabilidad de que el texto haya sido generado por un modelo de lenguaje bien entrenado.

7.3. *Métricas de Evaluación*

7.3.1. *BLEU (Bilingual Evaluation Understudy)*

Evalúa la similitud entre la respuesta generada y la respuesta esperada mediante el análisis de n-gramas, tomando en cuenta palabras y combinaciones de palabras. Para evitar puntajes nulos en respuestas cortas, se utilizó una función de suavizado.

7.3.2. ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

Mide la coincidencia entre la respuesta generada y la esperada considerando:

- ROUGE-1: Coincidencias de unigramas (palabras individuales).
- ROUGE-2: Coincidencias de bigramas.

- ROUGE-L: Longitud de la subsecuencia más larga común (LCS, por sus siglas en inglés).

7.3.3. *Perplejidad*

Se calculó procesando las respuestas generadas con el modelo GPT-2, obteniendo la exponencial de la pérdida (loss) como resultado.

7.4. *Procedimiento de Evaluación*

Para cada par de respuesta generada y esperada, se calcularon las métricas mencionadas siguiendo estos pasos:
Cálculo de BLEU

Se procesaron las respuestas generadas y esperadas como listas de palabras y se calculó la similitud usando la función sentence_bleu de la biblioteca NLTK con suavizado.
Cálculo de ROUGE

Con la biblioteca rouge_score, se calcularon las métricas ROUGE-1, ROUGE-2 y ROUGE-L considerando el balance entre precisión y exhaustividad mediante el f-measure.
Cálculo de Perplejidad

Se tokenizaron las respuestas generadas utilizando el tokenizador del modelo GPT-2. Posteriormente, se calcularon las probabilidades de las secuencias y se obtuvo la perplejidad como métrica final.

7.5. *Resultados*

Los resultados de las métricas (BLEU, ROUGE y Perplejidad) fueron añadidos al conjunto de datos original como nuevas columnas. Finalmente, los datos completos, que incluyen las preguntas, las respuestas generadas, las respuestas esperadas, los tiempos de respuesta y las métricas, se exportaron a una hoja de cálculo para su análisis. En la T

abla 2, se pueden observar los resultados de las métricas para cada una de las preguntas.

Tabla 2. Resultados de las métricas

Pregunta	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	Perplejidad
1	1	1	1	1	86.68
2	1	1	1	1	122.85
3	1	1	1	1	100.48
4	0.0563	0.571	0.315	0.571	463.04
5	0.599	0.732	0.724	0.732	66.76
6	1	1	1	1	110.50
7	0.013	0.137	0.074	0.137	164.52
8	1	1	1	1	133.99
9	1	1	1	1	80.09
10	1	1	1	1	126.99
11	1	1	1	1	275.00
12	1	1	1	1	39.08
13	1	1	1	1	73.24
14	1	1	1	1	156.86
15	1	1	1	1	304.38
16	1	1	1	1	149.96
17	1	1	1	1	124.68
18	1	1	1	1	320.03
19	1	1	1	1	88.11
20	1	1	1	1	562.64
21	1	1	1	1	96.70
22	1	1	1	1	84.02
23	1	1	1	1	75.35
24	1	1	1	1	62.36
25	1	1	1	1	82.93

7.6. *Importancia de las Métricas*

BLEU y ROUGE permiten evaluar la precisión semántica de las respuestas generadas, identificando la similitud entre las respuestas del sistema y las respuestas ideales, mientras que la perplejidad mide la calidad lingüística y la fluidez del texto generado, proporcionando una evaluación cuantitativa de la coherencia de las respuestas.

El análisis conjunto de estas métricas proporciona una visión integral del desempeño del sistema, permitiendo evaluar tanto la calidad semántica como la lingüística de las respuestas generadas. Este enfoque garantiza una evaluación objetiva y robusta, alineada con los estándares del procesamiento de lenguaje natural.

8. **Discusión**

Es importante destacar que la evaluación se realizó sobre una muestra de 25 preguntas representativas de los tipos de reglamentos: académico, administrativo, ético y de conducta, operativo, así como legal y normativo. Por lo que, no se consideró cubrir el total de los tipos de normativas existentes en la universidad. Los resultados muestran que en 22 de las 25 preguntas realizadas al sistema RAG, se generaron respuestas que presentan valores perfectos (1) de acuerdo con las métricas BLEU, ROUGE-1, ROUGE-2 y ROUGE-L, lo cual indica una alta similitud léxica y estructural con las respuestas esperadas. Esto sugiere que el sistema presentó una alta precisión en términos de contenido, ya que dichas métricas cuantifican la coincidencia entre la respuesta generada y la ideal previamente definida por expertos. No obstante, la perplejidad mostró una mayor variabilidad, lo que sugiere que, si bien las respuestas fueron precisas en contenido, la coherencia y fluidez del texto pueden no haber sido óptimas en todos los casos. Para futuras evaluaciones, se propone complementar el análisis con métricas cualitativas o evaluaciones humanas que permitan capturar con mayor precisión estos aspectos discursivos.

8.1. *Ejemplos de preguntas y resultados*

A continuación, se presentan algunos ejemplos representativos del desempeño del sistema, junto con sus respectivas métricas de evaluación.

8.1.1. Ejemplo 1. Pregunta 4: "¿Cuántas veces puedo presentar un examen extraordinario por semestre?"

La respuesta generada (“Puedes presentar un Examen Extraordinario hasta tres veces por semestre”) y la respuesta esperada (“Puedes presentar como máximo tres exámenes extraordinarios por semestre”) son semánticamente similares, ya que ambas comunican el mismo límite. No obstante, las métricas reflejan una baja coincidencia léxica y estructural como se puede observar en la Tabla 3.

Tabla 3. Métricas para la pregunta 4, ¿Cuántas veces puedo presentar un examen extraordinario por semestre?

BLEU	ROUGE-1	ROUGE-2	ROUGE-L	Perplejidad
0.056376	0.571	0.315	0.571	463.04

El valor bajo de BLEU indica diferencias significativas en la formulación textual, a pesar de la coincidencia conceptual. La perplejidad alta sugiere que el modelo tuvo dificultades para construir una respuesta fluida y natural. Este caso subraya la necesidad de ajustar el sistema para lograr no solo precisión semántica, sino también coherencia y estilo formal acorde al lenguaje de los reglamentos.

8.1.2. Ejemplo 2. Pregunta 7: "¿Cuántos créditos debo cursar para obtener mi título?"

La respuesta generada ("No se menciona en el texto proporcionado cuántos créditos debes cursar para obtener tu título") se aleja del enfoque de la respuesta esperada ("La cantidad de créditos que indique cada plan de estudios"), lo cual se refleja en métricas muy bajas como se muestra en la Tabla 4.

Tabla 4. Métricas para la pregunta 7: ¿Cuántos créditos debo cursar para obtener mi título?

BLEU	ROUGE-1	ROUGE-2	ROUGE-L	Perplejidad
0.013218	0.137	0.074	0.137	164.52

La baja similitud textual y semántica sugiere que el modelo no logró recuperar información relevante. Este resultado destaca la necesidad de fortalecer la capacidad del sistema para interpretar adecuadamente preguntas donde la información requerida no está explícitamente en el texto.

8.1.3. *Ejemplo 3. Pregunta: "¿Qué documentos necesito para inscribirme al servicio social?"*

La respuesta generada coincide exactamente con la respuesta esperada: "Acta de nacimiento, certificado de estudios, y otros documentos especificados en el reglamento." En este caso, las métricas reflejan una coincidencia total. Se muestra un extracto de los datos en la Tabla 5.

Este ejemplo muestra el potencial del sistema cuando la información es clara y accesible en los reglamentos. La baja perplejidad refuerza que la respuesta fue generada con fluidez y coherencia.

Tabla 5. Métricas para la pregunta 12: ¿Qué documentos necesito para solicitar mi título?

BLEU	ROUGE-1	ROUGE-2	ROUGE-L	Perplejidad
1	1	1	1	39.08

La respuesta generada coincide perfectamente con la respuesta esperada, lo que se refleja en valores perfectos de BLEU y ROUGE. Además, la perplejidad baja confirma que el modelo generó una respuesta coherente y fluida. Este es un ejemplo de un caso en el que el sistema funcionó de manera óptima, demostrando su capacidad para recuperar y generar información precisa y contextualizada.

Este caso resalta la eficacia del sistema cuando la información requerida está claramente especificada en los reglamentos y el modelo puede recuperarla sin ambigüedades. Sin embargo, también subraya la importancia de garantizar que los reglamentos estén bien estructurados y organizados para facilitar la recuperación de información. En general, este resultado positivo refuerza la viabilidad del sistema para proporcionar respuestas precisas y útiles en contextos donde la información es clara y accesible.

8.1.4. Ejemplo 4. Pregunta 11: "¿Cómo solicito una justificación de inasistencia?"

Aunque la respuesta generada es idéntica a la esperada ("Debe presentarse documentación justificativa en el tiempo estipulado"), la perplejidad fue elevada. El resumen de los datos para el análisis de la respuesta lo encontramos en la Tabla 6.

Tabla 6. Métricas para la pregunta 11: ¿Cómo solicito una justificación de inasistencia?

BLEU	ROUGE-1	ROUGE-2	ROUGE-L	Perplejidad
1	1	1	1	275.00

Este resultado indica que, aunque el contenido es correcto, el modelo no siempre genera texto con naturalidad, posiblemente debido a patrones de entrenamiento o estilo sintáctico.

Este caso resalta que, aunque el sistema es capaz de generar respuestas precisas y alineadas con la información esperada, aún existen áreas de mejora en cuanto a la coherencia y fluidez del texto generado. La alta perplejidad indica que el modelo podría beneficiarse de un ajuste adicional para mejorar la naturalidad del lenguaje, especialmente en respuestas que requieren un tono más formal o técnico. En general, este resultado demuestra la eficacia del sistema para proporcionar respuestas precisas, pero también subraya la importancia de trabajar en la calidad lingüística de las respuestas generadas.

En general, el sistema es eficiente en la generación de respuestas precisas, pero se necesitan ajustes para mejorar la coherencia y fluidez del texto generado, especialmente en casos donde la información no está claramente especificada en los reglamentos.

9. Conclusiones

El sistema propuesto demuestra ser una herramienta eficaz para mejorar el acceso a la información normativa en organismos públicos. La utilización de RAG y modelos de lenguaje preentrenados permite generar respuestas precisas y contextualizadas en tiempo real, lo que optimiza el tiempo de consulta y mejora la eficiencia operativa. Aunque el sistema genera respuestas precisas en términos de contenido, se observa que la coherencia y fluidez del texto generado pueden mejorar, especialmente en casos donde la información no está claramente especificada en los reglamentos. Esto sugiere que el modelo podría beneficiarse de un ajuste adicional para manejar consultas ambiguas o complejas.

Las métricas utilizadas (BLEU, ROUGE y Perplejidad) proporcionan una evaluación integral del sistema, permitiendo analizar tanto la precisión semántica como la calidad lingüística de las respuestas generadas. Esto asegura que el

sistema cumpla con los estándares de calidad en PLN. El sistema tiene un alto potencial de aplicación en diferentes contextos públicos, como administraciones gubernamentales, instituciones educativas y otros organismos que manejen un gran volumen de normativas. La capacidad de personalización y escalabilidad del sistema lo hace adaptable a las necesidades específicas de cada organización.

Aunque el sistema es preciso en términos de contenido, la coherencia y fluidez del texto generado puede mejorar, especialmente en respuestas que requieren un tono más formal o técnico. El sistema podría mejorar su capacidad para manejar consultas ambiguas o complejas, donde la información no está explícitamente mencionada en los reglamentos. Una mejor indexación de los reglamentos podría facilitar la recuperación de información específica, mejorando la precisión de las respuestas generadas.

Referencias

Administración Electrónica. (2022). Chatbots o asistentes virtuales en las administraciones públicas para democratizar el uso de datos abiertos. Recuperado de https://administracionelectronica.gob.es

Cohere. (2024). Empowering AI with Language Models. Recuperado de https://cohere.ai

Copilot.Live. (2024). Chatbots especializados para servicios gubernamentales. Recuperado de https://www.copilot.live

DeepSeek. (2024). *DeepSeek: Advanced NLP and RAG integration*. Recuperado de https://deepseek.com

El Universal. (2024). Con TEO, un chat virtual, buscan combatir mal servicio y corrupción dentro de la Fiscalía de Oaxaca. El Universal Oaxaca. Recuperado de https://oaxaca.eluniversal.com.mx/estatal/con-teo-un-chat-virtual-buscan-combatir-mal-servicio-y-corrupcion-dentro-de-la-fiscalia-de

Facebook AI. (2021). FAISS: A Library for Efficient Similarity Search. Recuperado de https://github.com/facebookresearch/faiss

Gheorghiu, A. (2024). Building Data-Driven Applications with LlamaIndex: A Practical Guide to Retrieval-Augmented Generation (RAG) to Enhance LLM Applications. Packt Publishing.

Gheorghiu, M. (2024). LlamaIndex: Revolutionizing Information Retrieval with Large Language Models. Recuperado de https://llamaindex.ai

Gobierno de la Ciudad de Buenos Aires. (2024). Caso Boti: Chatbot para la Ciudad de Buenos Aires. Recuperado de https://buenosaires.gob.ar

Greyling, C. (2023). *Agents & LLMs: Multihop Question Answering*. Medium. Recuperado de https://cobusgreyling.medium.com/agents-llms-multihop-question-answering-ca6521227b6c

Hugging Face. (2022). Transformers for Multilingual Applications. Recuperado de https://huggingface.co

LangChain. (2025). *LangChain Documentation*. Recuperado de https://www.langchain.com/

Laparra, E. (2024). Retrieval-Augmented Generation para la extracción de información de documentos inteligentes. Recuperado de https://m.riunet.upv.es

LlamaIndex. (2024). *LlamaIndex: Documentation*. Recuperado de https://docs.llamaindex.ai/

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

OpenAI. (2023). GPT-4: Capabilities and Use Cases. Recuperado de https://openai.com/research/gpt-4

Patel, A. (2023). LLM Evaluation Metrics Explained. Medium. Recuperado de https://medium.com/data-science-in-your-pocket/llm-evaluation-metrics-explained-af14f26536d2

Sentence-Transformers. (2024). paraphrase-multilingual-MiniLM-L12-v2. Hugging Face. Recuperado de https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

Streamlit. (2024). Streamlit Documentation. Recuperado de https://docs.streamlit.io/

Takahashi, S., & Tanaka-Ishii, K. (2019). *Evaluating computational language models with scaling properties of natural language*. *Computational Linguistics*, 45(3), 481–513. https://doi.org/10.1162/coli_a_00355

Universidad Tecnológica de la Mixteca. (n.d.). Legislación universitaria. Recuperado el 11 de noviembre de 2024, de https://www.utm.mx/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.