Revista Cubana de Ciencias Informáticas
Vol. X, No. Y, Mes Xxx-Xxx, 20xx
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. x-y

Tipo de artículo: Artículo originales
Temática: Inteligencia artificial
Recibido: dd/mm/aa | Aceptado: dd/mm/aa | Publicado: dd/mm/aa

Democratización de Herramientas de Inteligencia Artificial para la Gestión de Documentos PDF en Contextos con Recursos Limitados

Democratization of Artificial Intelligence Tools for PDF Document Management in Resource-Limited Contexts

Joaquin Rivas Sánchez 0009-0007-7971-8259[1*]
Naylin Brizuela Capote 0009-0000-5059-8248[1]
Angel Alberto Vazquez Sánchez 0000-0002-3130-7983[1]

[1]Universidad de las Ciencias Informáticas. Km 2 ½ carretera a San Antonio de los Baños, Reparto Torrens, La Lisa, La Habana
[*]Autor para correspondencia: (joaquiners@estudiantes.uci.cu)

RESUMEN

Este artículo abordó la necesidad de desarrollar una herramienta accesible para la gestión y análisis de documentos PDF en entornos con recursos limitados, como el contexto cubano, donde las restricciones de conectividad y métodos de pago internacionales dificultan el acceso a soluciones avanzadas como ChatPDF y Humata.ai. Además, se consideró crucial avanzar en el desarrollo de tecnologías que promuevan la soberanía tecnológica, reduciendo la dependencia de plataformas extranjeras. Para solucionar este problema, se propuso un prototipo basado en la técnica Retrieval-Augmented Generation (RAG), empleando el modelo de lenguaje llama-3.2-3B y all-mpnet-base-v2 para la generación de word embeddings. La principal funcionalidad del prototipo incluye la capacidad de responder preguntas abiertas (open-domain question answering). Adicionalmente, el prototipo demostró su potencial para funcionar eficazmente en entornos con bajos recursos. Se concluye que este prototipo puede ser desarrollado aún más y tiene el potencial de democratizar el acceso a tecnologías avanzadas de inteligencia artificial, lo que fortalecería la capacidad local para gestionar documentos PDF de manera eficiente.

Palabras clave: inteligencia artificial; RAG; procesamiento de lenguaje natural; conectividad limitada; código abierto.

Revista Cubana de Ciencias Informáticas
Vol. X, No. Y, Mes Xxx-Xxx, 20xx
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. x-y

ABSTRACT

This article addressed the need to develop an accessible tool for managing and analyzing PDF documents in resource-limited environments, such as the Cuban context, where connectivity restrictions and international payment methods hinder access to advanced solutions like ChatPDF and Humata.ai. Furthermore, advancing the development of technologies that promote technological sovereignty and reduce reliance on foreign platforms was deemed crucial. To address this issue, a prototype was proposed based on the Retrieval-Augmented Generation (RAG) technique, utilizing the llama-3.2-3B language model and all-mpnet-base-v2 for generating word embeddings. The prototype's main functionality includes the ability to perform open-domain question answering. Additionally, the prototype demonstrated its potential to function effectively in low-resource settings. It is concluded that, with further development, this prototype has the potential to democratize access to advanced artificial intelligence technologies, strengthening local capabilities for efficient PDF document management.

Keywords: artificial intelligence; RAG; natural language processing; limited connectivity; open source.

Introducción

En la actualidad, gran parte del conocimiento colectivo se encuentra predominantemente digitalizado manifestándose en diversas formas como noticias, blogs, literatura científica y documentos empresariales. Este fenómeno ha sido impulsado por la revolución digital y el uso extensivo de las redes sociales, que han transformado la manera en que las personas interactúan y comparten información. En estos contextos, los documentos PDF representan una fuente crítica de información (Delgado López, 2022), y el acceso eficiente a herramientas avanzadas para su análisis puede marcar una diferencia significativa en la productividad y la comprensión del contenido. En países como Cuba, los profesionales enfrentan limitaciones significativas que afectan su capacidad para acceder a herramientas tecnológicas avanzadas, como plataformas de análisis basadas en inteligencia artificial (IA) como ChatPDF, Humata.ai o ChatGPT. Estas limitaciones incluyen conectividad limitada y barreras de acceso a métodos de pago internacionales, lo que restringe el uso efectivo de estas herramientas que podrían potenciar la gestión y análisis de documentos PDF. Estas plataformas, además de requerir una conexión constante a internet, suelen operar mediante modelos de pago que exigen

tarjetas de crédito o suscripciones en moneda extranjera, a los cuales muchos profesionales cubanos no tienen acceso. Esta situación crea un obstáculo en el aprovechamiento de herramientas que, aunque altamente útiles, resultan inviables en entornos con restricciones de recursos tecnológicos y financieros.

La presente investigación tiene como objetivo adaptar y replicar soluciones existentes, para ofrecer una opción que permita a los usuarios cubanos beneficiarse de las ventajas de la IA en la gestión de documentos PDF. Esto refleja un enfoque de innovación, que no siempre implica investigación y desarrollo, sino la aplicación de conocimientos actuales para resolver problemas de manera eficiente (Huang and Huang, 2024). La técnica RAG aborda desafíos como la alucinación y el conocimiento desactualizado en los Modelos de Lenguaje de Gran Escala (LLMs), mejorando la exactitud y la fiabilidad en tareas como la respuesta a consultas, la resumen de documentos y la verificación de hechos. Este enfoque permite que los LLMs accedan a información actualizada y relevante, superando las limitaciones de los modelos tradicionales que dependen únicamente de datos estáticos (Gao et al., 2024). En el campo de la gestión de documentos y procesamiento de lenguaje natural, existen múltiples investigaciones que respaldan el uso de RAG. Este trabajo se basa en la revisión de estudios que describen el estado actual de esta técnica, con el objetivo de adaptarla a las condiciones específicas de conectividad y recursos de Cuba. La propuesta se enfoca en desarrollar un prototipo de herramienta que, aunque emplea tecnologías ya probadas, esté diseñada para funcionar localmente en entornos con infraestructura limitada, sin depender de servidores en la nube ni requerir pagos internacionales.

Los objetivos de este trabajo son:

1. Identificar las limitaciones actuales en el uso de herramientas de IA para la gestión de documentos PDF en el contexto cubano.

2. Adaptar y replicar una solución basada en la técnica RAG para ofrecer una alternativa que considere las condiciones de conectividad así como las limitaciones de infraestructura de nuestro país.

3. Evaluar el desempeño de la herramienta propuesta en un entorno con limitaciones tecnológicas, con énfasis en su viabilidad y utilidad práctica.

4. Proponer futuras líneas de desarrollo para extender la funcionalidad de la herramienta, considerando la integración con otras tecnologías emergentes. Esto incluirá recomendaciones para mejorar la capacitación en el uso de herramientas digitales y fomentar un ecosistema que apoye la innovación en la gestión documental.

Desarrollo

Tabla 1 - Comparación de Productos Similares.

Herramienta	Bloqueada en Cuba	Límites en Planes Gratuitos
ChatPDF	X	10 MB por archivo; hasta 120 páginas gratis; 20 respuestas al dia
Humata.ai	X	60 páginas; 10 respuestas
ChatGPT	X	3 documentos diarios / 7 respuestas
UPDF	X	5 documentos - 100 páginas

Estas herramientas, aunque útiles, no están diseñadas para contextos como el cubano, donde la conectividad a internet es intermitente y no existe acceso a métodos de pago internacionales. Además, el acceso es limitado, ya que, por ejemplo, ChatPDF permite ingresar pero no procesar ningún documento sin el uso de una Red Privada Virtual (VPN) para evitar las restricciones de acceso desde Cuba, lo que las hace inútiles en el uso diario de investigadores cubanos.

Evaluación de soluciones

La solución propuesta en este trabajo se basa en una revisión de las técnicas disponibles para extender las capacidades de los LLMs mediante la integración de nuevas fuentes de información. Las principales soluciones evaluadas fueron el fine-tuning (FT) y el Retrieval-Augmented Generation (RAG), ambas con características distintivas que influyen en su aplicabilidad en diferentes contextos. Tras un análisis detallado, se propone RAG como la técnica central para este trabajo, dado su mejor ajuste a las necesidades del proyecto y su capacidad para operar en entornos con limitaciones de infraestructura, ver Tabla 2.

La técnica Retrieval-Augmented Generation (RAG) fue seleccionada para este trabajo debido a varias razones que la hacen ideal para el contexto cubano. Al ser una técnica que mejora las capacidades de los modelos de lenguaje sin necesidad de reajustar los modelos base y ser más flexible a cambios constantes en los datos, es fundamental en entornos con recursos computacionales limitados (Muludi et al., 2024). La capacidad de RAG para reducir el tamaño del modelo y soportar contextos más largos resulta especialmente ventajosa en aplicaciones como los chatbots, que a menudo requieren manejar conversaciones largas y complejas. Además, su flexibilidad para incorporar y actualizar conocimientos externos, incluidos datos especializados y confidenciales, es un recurso valioso para chatbots que necesitan proporcionar información relevante y actualizada a los usuarios (Gao et al., 2024).

Tabla 2 - Comparativa entre Fine-Tuning y RAG.

Características	**Fine-tuning (FT)**	**Retrieval-Augmented Generation (RAG)**
Adaptación del Modelo	Ajusta un modelo preentrenado sin necesidad de reentrenarlo completamente para cada nueva actualización de conocimiento	No requiere reentrenamiento o reajuste; permite agregar conocimiento externo en tiempo real.
Requerimientos Computacionales	Altos: necesita grandes cantidades de datos y recursos computacionales para entrenar.	Menores en comparación, ya que se enfoca en la adaptación, integración y recuperación de la información
Capacidades Dinámicas	Menos adaptable a entornos cambiantes; el modelo es estático tras el entrenamiento.	Altamente dinámico; permite integrar nueva información continuamente.
Latencia	Baja, ya que el modelo responde directamente con el conocimiento integrado	Mayor, debido al proceso de recuperación y generación de respuestas en tiempo real
Aplicaciones	Ideal para tareas que requieren replicar estructuras o estilos específicos	Perfecto para tareas de recuperación de información y generación basada en fuentes externas
Limitaciones	No adecuado para incorporar rápidamente nuevos conocimientos, posibles preocupaciones éticas sobre la recuperación de datos	Mayor complejidad técnica y, debido a tener más partes móviles, se incrementa el riesgo de un único punto de falla

Métodos o Metodología Computacional

Para el desarrollo del prototipo se utilizó Python v3.12, seleccionado por su versatilidad y compatibilidad con bibliotecas de inteligencia artificial y procesamiento de texto. Entre las principales bibliotecas empleadas se encuentran Giskard v2.15.2, utilizada para la evaluación del sistema; Faiss-cpu v1.9.0, que permitió realizar búsquedas vectoriales rápidas y precisas; y Gradio v5.3.0, utilizada para construir una interfaz gráfica que facilitó la interacción y evaluación del sistema. Además, se integró llama-cpp v0.2.90 para cargar y ejecutar localmente el modelo Llama 3.2-3B, y PyTorch v2.4.1+cu124, que proporcionó el soporte necesario para las operaciones de aprendizaje profundo, optimizando el rendimiento en diferentes entornos. Como se concluyó en la sección anterior, la solución seleccionada fue Retrieval-Augmented Generation (RAG), que combina la búsqueda de información con generación de respuestas. En este contexto, se integró el modelo de lenguaje preentrenado Llama-3.2-3B-Instruct, encargado de las tareas de comprensión y generación de texto, y el modelo all-mpnet-base-v2 para la generación de embeddings, con el fin de optimizar el procesamiento semántico de los documentos.

Las pruebas iniciales y las fases tempranas de desarrollo se llevaron a cabo en el entorno de computación en la nube Google Colab, debido a su facilidad de configuración y acceso a recursos de GPU. Sin embargo, las fases posteriores del desarrollo se ejecutaron localmente, utilizando un equipo con un procesador Intel Core i7 7700k, 16 GB de RAM y una tarjeta gráfica NVIDIA RTX 2070 en el sistema operativo Arch Linux, lo que permitió ejecutar las tareas de procesamiento y generación de manera eficiente, sin presentar limitaciones de

hardware.

Metodología de Pruebas

La validación del prototipo se llevó a cabo utilizando tres artículos científicos de distintas áreas del conocimiento:

1. "Attention is All You Need": Introducción de los transformers, un modelo de inteligencia artificial desarrollado por Google (Vaswani et al., 2023).
2. "Estudio cubano sobre vacunación contra COVID-19": Análisis de la seguridad y eficacia de dos esquemas vacunales (Toledo-Romaní et al., 2023).
3. "Reporte del IPCC 2023": Documento clave sobre tendencias de cambio climático y medio ambiente (Lee et al., 2023).

Con estos se diseñaron tres conjuntos de prueba de 20 preguntas cada uno, abarcando diferentes aspectos clave: resistencia a preguntas distractoras, capacidad de desambiguación en preguntas dobles, ignorancia de contexto irrelevante y manejo del contexto conversacional previo. La biblioteca Giskard fue utilizada para generar estas preguntas y realizar la evaluación del prototipo, la cual incluye el toolkit RAGET (Retrieval-Augmented Generation Evaluation Toolkit), diseñado para evaluar los componentes clave de un sistema RAG:

- Retriever: Encargado de recuperar información relevante del conjunto de documentos.
- Generator: Utiliza un LLM para generar respuestas basadas en los contextos recuperados.
- Rewriter: Reformula consultas para mejorar su relevancia o adaptarlas al contexto previo.
- Router: Filtra consultas según las intenciones del usuario, optimizando la interacción.
- Knowledge Base: Base de conocimiento que almacena los documentos utilizados para generar respuestas.

Giskard evalúa el sistema mediante métricas específicas para cada componente

- Recall Contextual: Valora la eficacia del retriever en encapsular la información necesaria para las respuestas.
- Relevancia de la Respuesta: Mide qué tan pertinentes y precisas son las respuestas generadas en relación con las consultas y contextos recuperados.
- Fidelidad: Examina si las respuestas están fundamentadas en los documentos recuperados, evitando errores factuales.

En esta validación, se utilizó la métrica Accuracy, que mide la proporción de respuestas correctas proporcionadas por el sistema frente a las respuestas esperadas (Gao et al., 2024).

Adicionalmente, se realizaron pruebas similares a dos sistemas homólogos: ChatPDF y Humata.ai. Estas pruebas se llevaron a cabo con el objetivo de evaluar las capacidades de estos sistemas en tareas específicas, como la extracción y comprensión de información en documentos PDF. El análisis comparativo resultante no sólo establece un punto de referencia claro para medir el desempeño del prototipo desarrollado, sino que también permite identificar áreas de mejora y optimización. Este enfoque contribuye a orientar futuras iteraciones del sistema, garantizando que las soluciones propuestas sean competitivas y alineadas con los estándares de desempeño actuales en la industria.

Implementación y desarrollo de RAG

El enfoque de Retrieval-Augmented Generation (RAG) en este prototipo sigue la estructura tradicional, que se divide en tres fases principales: indexado de documentos, recuperación y generación aumentada. Esta metodología optimiza el rendimiento de los LLMs al combinar sus capacidades generativas con la recuperación de información desde memoria no paramétrica, lo que permite mejorar significativamente la precisión y relevancia de las respuestas y reducir así las alucinaciones (Zhang and Kotanko, 2024).

A continuación, se describe la estructura de RAG implementada en el prototipo, destacando los pasos clave y las optimizaciones realizadas para su adaptación a un entorno con limitaciones de hardware e infraestructura como el cubano.

Fases de RAG en el Prototipo

1. Indexado de Documentos: En esta fase inicial, los documentos PDF se fragmentan en çhunks.º porciones más pequeñas de texto. Estos chunks se transforman en vectores de embeddings, que son representaciones numéricas de palabras, párrafos o documentos en un espacio de alta dimensión (Dai et al., 2015). El sistema utiliza el modelo pre-entrenado all-mpnet-base-v2, para generar estos embeddings. Posteriormente, los vectores se almacenan en una base de datos de vectores, optimizada mediante tecnologías como Facebook AI Similarity Search (FAISS), una biblioteca de código abierto desarrollada por Facebook AI Research para la búsqueda de similitudes de forma eficiente y el agrupamiento de vectores densos (Douze et al., 2024). En el contexto de este prototipo se incluyen metadatos relevantes (como temas o secciones del documento) para mejorar la precisión de la recuperación. Esto es esencial para asegurar que los chunks relevantes se identifiquen rápidamente, incluso en un corpus extenso y variado.

2. Recuperación de Información: En esta segunda fase, se realiza la búsqueda de los k ?chunks? más relevantes que pueden responder a una consulta o pregunta planteada por el usuario. El proceso de recuperación se basa en la búsqueda semántica, que encuentra información relacionada usando el contexto y la semántica de la pregunta. Para aumentar la precisión, se empleó un enfoque híbrido que combina búsqueda semántica con algoritmos como BM25 (un modelo de recuperación basado en palabras clave), logrando así una recuperación más robusta y precisa (Omrani et al., 2024). Se aplicaron técnicas de reescritura y expansión de la consulta para asegurar que las preguntas formuladas por los usuarios sean más claras y estén mejor adaptadas a la tarea de recuperación. Esto mejora la precisión de la búsqueda, maximizando la relevancia de los chunks recuperados (Chan et al., 2024).

3. Generación Aumentada: En la fase final, se crea un prompt que combina la pregunta inicial con los chunks más relevantes recuperados en la fase anterior. Este prompt se pasa al modelo de lenguaje, Llama 3.2 3B, que genera una respuesta coherente y precisa basada en la información recuperada. Esta etapa es crítica, ya que la generación de respuestas no solo se basa en el conocimiento integrado del modelo, sino también en la información actualizada y específica proporcionada por los documentos recuperados. Para mejorar la precisión de las respuestas y reducir la sobrecarga de información irrelevante, se implementaron técnicas de reordenamiento y compresión del contexto. Estas técnicas aseguran que solo la información más relevante se incorpore al prompt final, facilitando así una integración fluida entre el contenido recuperado y la consulta original (Shi et al., 2024)

Revista Cubana de Ciencias Informáticas
Vol. X, No. Y, Mes Xxx-Xxx, 20xx
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. x-y

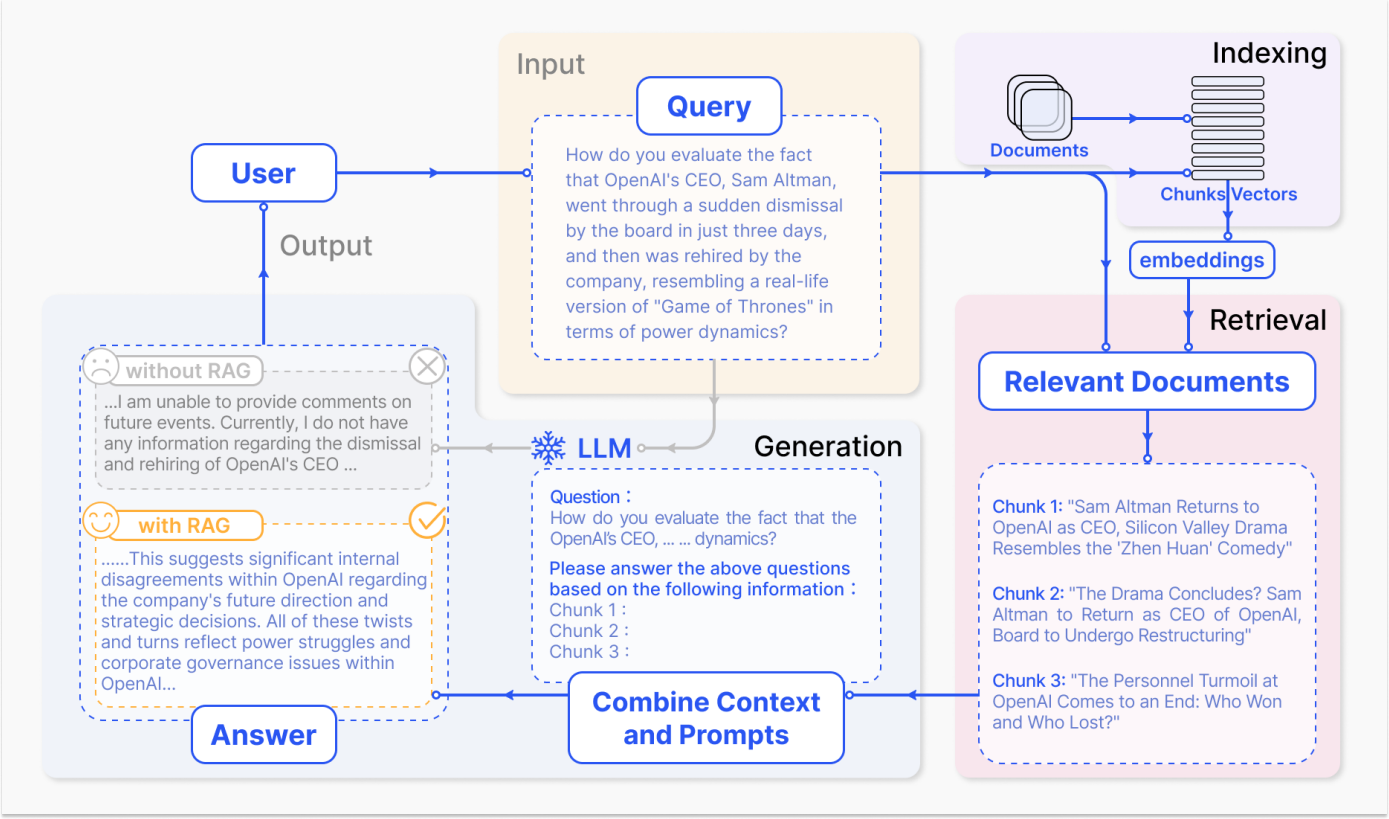

Fig. 1 - Arquitectura de un naive RAG Fuente: (Gao et al., 2024)

Diferentes niveles de complejidad

Los enfoques de RAG pueden clasificarse en tres tipos principales, cada uno con diferentes niveles de complejidad y optimización:

1. Naive RAG: Esta versión básica incluye las tres fases fundamentales: indexación, recuperación y generación. Aunque es eficiente en términos de implementación, tiene limitaciones en cuanto a la precisión y relevancia, debido a la falta de optimización en las fases de indexación y recuperación post-consulta (ver Fig. 1).

2. Advanced RAG: Este enfoque introduce optimizaciones adicionales, como el uso de ventanas deslizantes para mejorar la segmentación de documentos y la compresión del contexto para reducir la sobrecarga

de información innecesaria. También emplea técnicas avanzadas de reordenamiento y segmentación precisa para asegurar que la información más relevante se priorice (ver Fig. 2)

3. Modular RAG: Este es el enfoque más avanzado y flexible, que permite la introducción de múltiples módulos optimizados para diferentes fases del proceso (indexación, recuperación, generación y reordenamiento). Aunque ofrece una mejora considerable en términos de personalización y precisión, requiere una mayor capacidad computacional y puede resultar complicado de implementar en entornos con recursos limitados (ver Fig. 2)

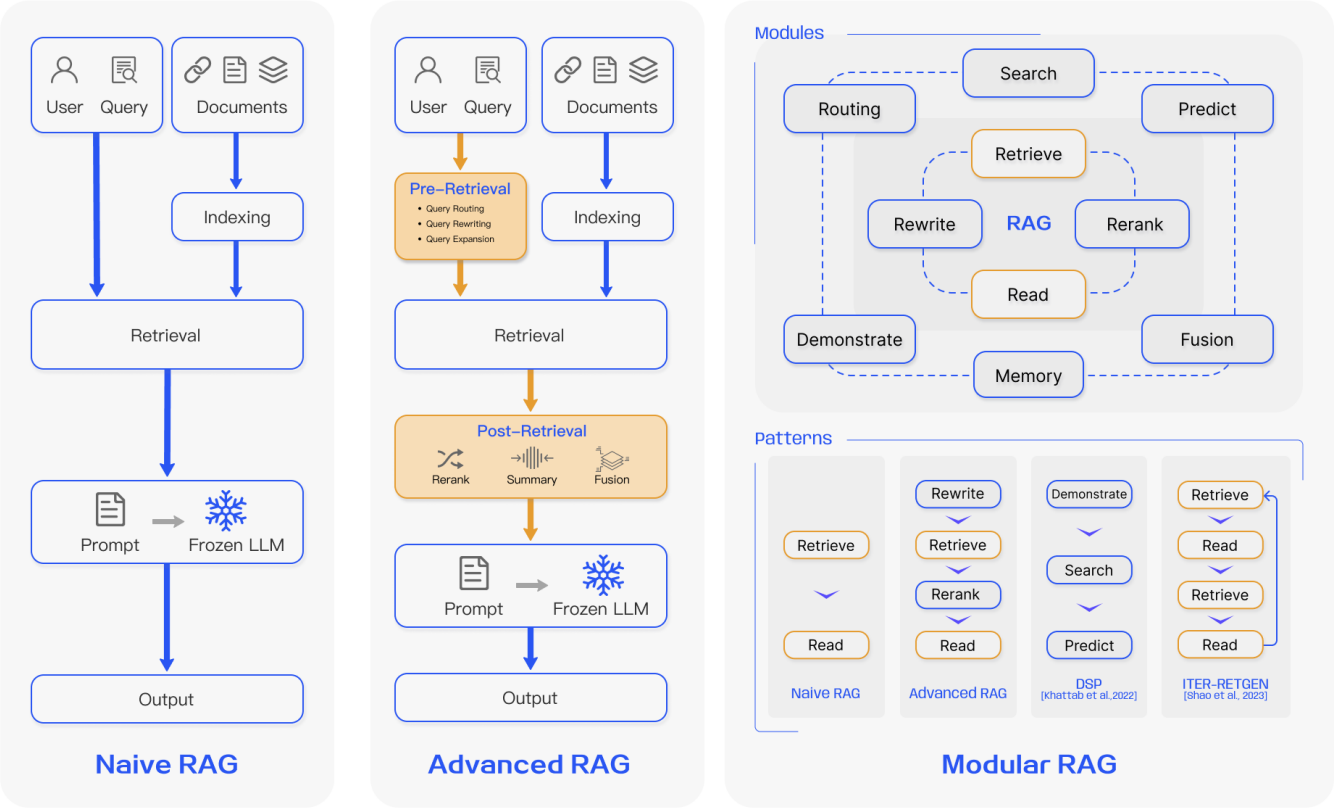

Fig. 2 - Diferentes tipos de RAG Fuente: (Gao et al., 2024)

Para el prototipo se seleccionó la arquitectura de Advanced RAG. La arquitectura avanzada de RAG equilibra precisión y recursos al mejorar la relevancia de la información recuperada mediante técnicas de segmentación

optimizada y reescritura de consultas, sin imponer una carga computacional excesiva, algo esencial en entornos con limitaciones de hardware como el cubano. Además, optimiza la generación de respuestas al reducir la inclusión de contenido irrelevante, utilizando estrategias de compresión y reordenamiento de contexto para asegurar coherencia y eficiencia (Gao et al., 2024). La propuesta de solución es la siguiente (ver Fig. 3)

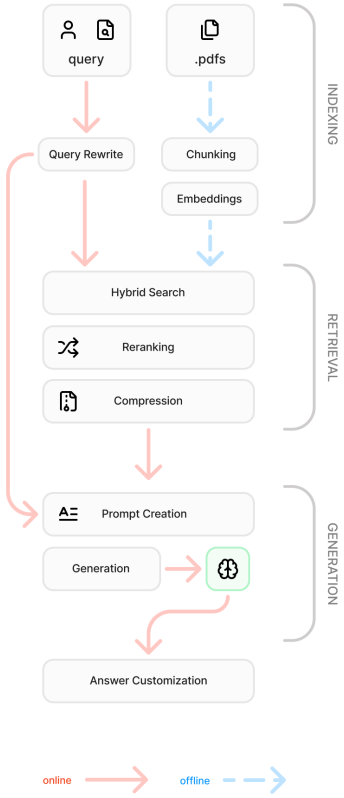

Fig. 3 - Propuesta de sistema

Resultados y discusión

El prototipo desarrollado fue evaluado en función de su capacidad para procesar documentos PDF y generar respuestas coherentes, manteniendo tiempos de respuesta rápidos. Los resultados obtenidos indican que,

aunque no se diseñó como un reemplazo de herramientas avanzadas como ChatPDF o Humata.ai, el prototipo demuestra un potencial significativo como una solución eficiente y adaptable, particularmente en contextos de recursos limitados. Debido a las limitaciones de recursos computacionales, las pruebas se limitaron a un pequeño conjunto de datos de prueba, lo que limita la exhaustividad del análisis.

Los resultados, presentados en la Tabla 3 y Fig 4, muestran diferencias notables entre el prototipo, ChatPDF y Humata.ai en los tres conjuntos de datos probados.

Tabla 3 - Resultados porcentuales.

Dataset	Sistema	Generator	Retriever	Rewriter	Routing	Knowledge Base	Overall Correctness
ipcc	Prototipo	40 %	40 %	40 %	100 %	0 %	40 %
	ChatPDF	85 %	80 %	80 %	100 %	0 %	85 %
	Humata	45 %	20 %	0 %	100 %	0 %	45 %
soberana	Prototipo	45 %	62.5 %	16.67 %	100 %	0 %	45 %
	ChatPDF	35 %	50 %	16.67 %	100 %	33.33 %	40 %
	Humata	45 %	62.5 %	25 %	100 %	0 %	40 %
attention	Prototipo	50 %	62.5 %	33.33 %	100 %	0 %	45 %
	ChatPDF	70 %	37.5 %	58.33 %	100 %	25.0 %	65 %
	Humata	40 %	37.5 %	16.67 %	100 %	50 %	35 %

[a] De emplear notas aclaratorias se colocarán al pie de la tabla.
[b] Otra nota aclaratoria.

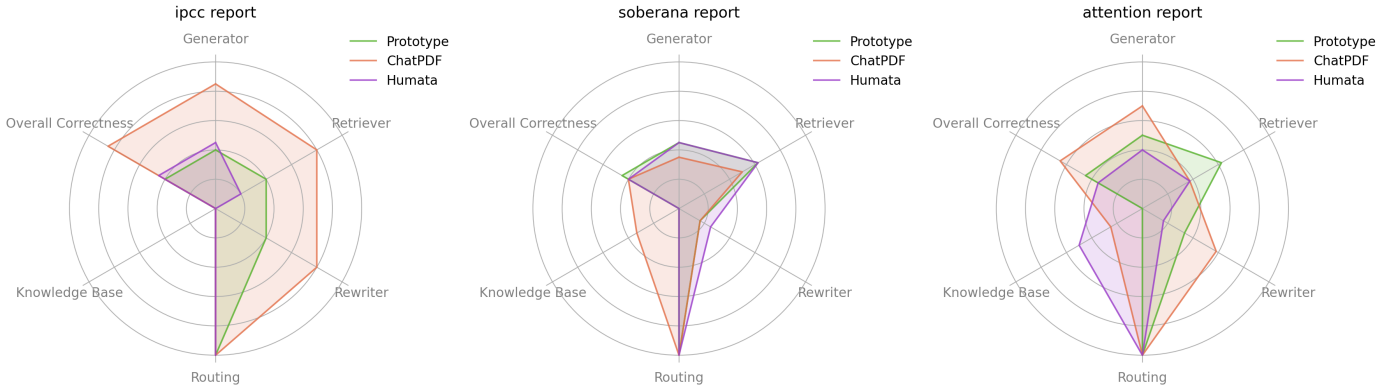

Fig. 4 - Gráfico comparativo

Basado en los resultados obtenidos, se pueden extraer las siguientes conclusiones:

1. El prototipo muestra un desempeño sólido en varias áreas, con especial fortaleza en la generación de respuestas. Sin embargo, su rendimiento es menor al de ChatPDF y Humata en ciertos aspectos, lo que

podría deberse a desafíos en el procesamiento de gráficos, tablas y fórmulas, como los presentes en el dataset attention. Por otro lado, en documentos más textuales como soberana e ipcc, el sistema logró mayor consistencia en recuperación y reescritura. Estos hallazgos subrayan la importancia de mejorar el manejo de formatos más complejos para ampliar su aplicabilidad y eficacia.

2. Destaca en la recuperación de la base de conocimiento y en la calidad de las respuestas generadas, lo que evidencia un diseño sólido tanto del pipeline como del modelo generador.

3. El enrutamiento de la consulta y la reescritura de preguntas presentan oportunidades de optimización, especialmente en el dataset attention donde ChatPDF tiene una ventaja clara.

En conclusión, los resultados sugieren que el prototipo tiene un buen potencial para ser una alternativa localizada, con un rendimiento sólido en tareas relacionadas con la gestión de documentos PDF en contextos con recursos limitados, aunque requiere mejoras puntuales en componentes específicos para alcanzar o superar a herramientas comerciales como ChatPDF y Humata.

Un aspecto clave para futuras líneas de desarrollo es la implementación de modelos más pequeños, como Llama 3.2-1B creado para ser usado en teléfonos móviles. Estos modelos permitirían una mayor concurrencia de usuarios y facilitarían su ejecución en dispositivos del lado del cliente. Esto no solo reduciría la carga sobre los servidores, sino que también mejoraría la accesibilidad, especialmente en infraestructuras con recursos limitados (Meta, 2024). Además, se podría explorar la incorporación de un modelo multimodal, capaz de procesar tanto texto como imágenes, lo que ampliará la capacidad de comprensión a gráficos, diagramas y, potencialmente, formatos más allá del PDF (Parodi and Julio, 2017).

Una dirección interesante sería el uso combinado de distintos modelos en el sistema. Por ejemplo, emplear un modelo más potente para la generación de respuestas, mientras se utiliza uno más rápido y liviano para tareas como recuperación de información y reescritura. Este enfoque permitiría equilibrar rendimiento y eficiencia. Asimismo, es fundamental mejorar el preprocesamiento de documentos, ya que el rendimiento de todos los componentes de un sistema RAG depende en gran medida de un componente de indexado robusto. Una implementación deficiente podría degradar la recuperación de información crítica y, en consecuencia, la calidad general del sistema (Setty et al., 2024). En particular, el preprocesamiento debería extenderse a elementos como tablas, imágenes, fórmulas matemáticas y gráficos, haciendo uso de modelos multimodales avanzados, que podrían interpretar estos elementos de manera más efectiva (Han et al., 2023).

Estas mejoras no solo incrementarían la precisión y la versatilidad del prototipo, sino que también lo posicionaría como una solución más robusta y adaptable para la gestión de información en contextos desafiantes.

El potencial de expansión del prototipo es significativo. Podría implementarse en plataformas nacionales de aprendizaje como el Entorno Virtual de Aprendizaje, donde actuaría como un recurso de apoyo para los estudiantes, ayudándoles a entender mejor los materiales de estudio, especialmente en contextos con disponibilidad limitada de profesores. Además, su aplicación podría ampliarse al apoyo a investigadores cubanos, ofreciendo herramientas avanzadas para procesar el contenido de artículos científicos, generar resúmenes claros y concisos, traducir textos especializados con alta fidelidad, y extraer información relevante de documentos PDF. Esto no solo optimizaría su acceso al conocimiento global, sino que también fortalecería la producción científica al facilitar el análisis y manejo de grandes volúmenes de datos de manera eficiente y precisa. Asimismo, podría utilizarse en la atención de quejas empresariales, automatizando el procesamiento de documentos tanto de consumidores como de la empresa, y proporcionando respuestas precisas a sus inquietudes. Estos ejemplos destacan la versatilidad del prototipo y su capacidad para impactar positivamente en diversos campos.

Para complementar los resultados descritos, se presenta una captura del prototipo en funcionamiento, ver Fig. 5. En esta imagen, se puede observar la interfaz de usuario y el flujo de procesamiento de una consulta basada en un documento PDF, lo que ilustra las capacidades clave del sistema.

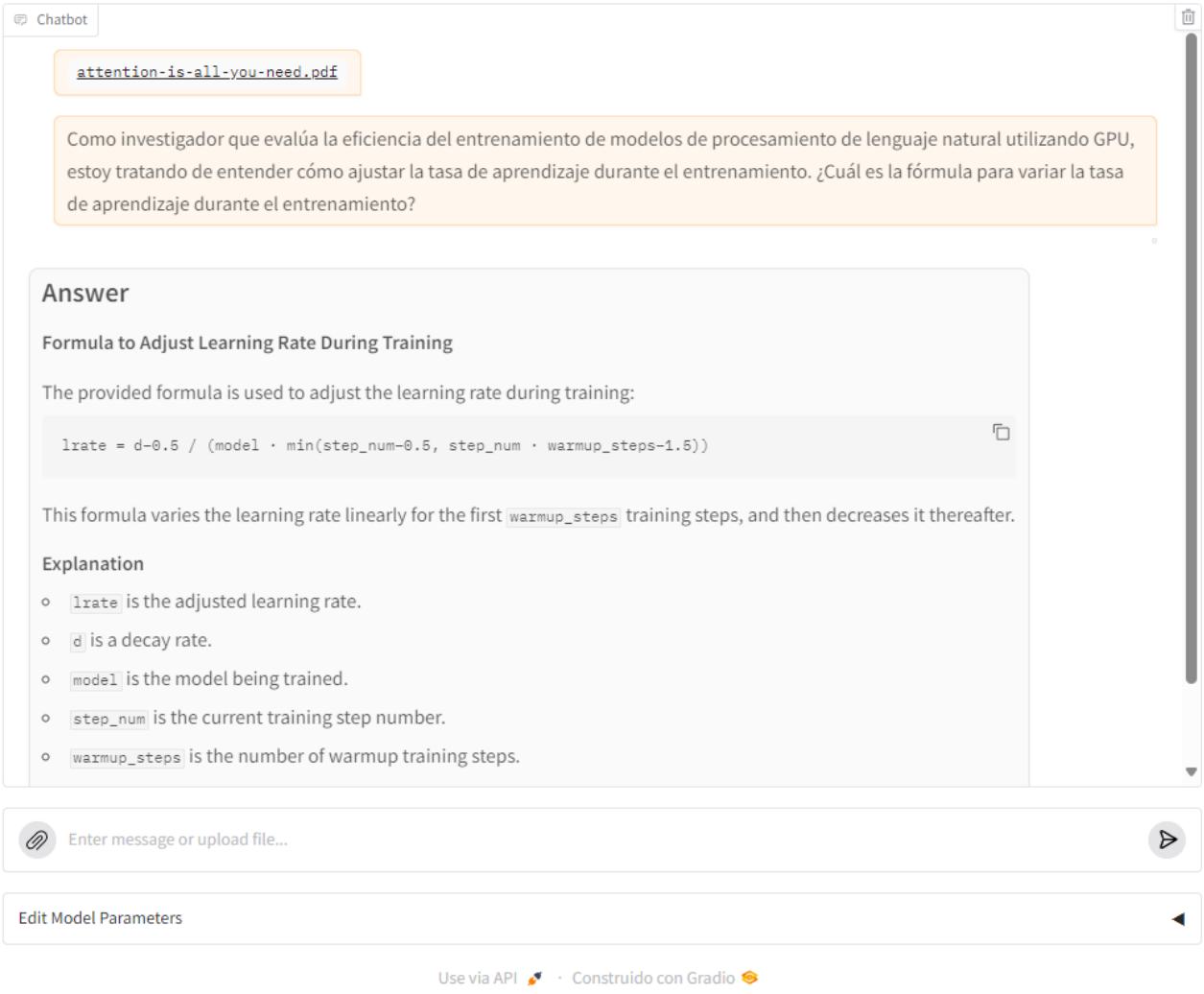

Fig. 5 - Interfaz del prototipo en funcionamiento procesando una consulta basada en un documento PDF.

Además, con el objetivo de fomentar la transparencia y reproducibilidad de los resultados, el código fuente del prototipo, junto con los conjuntos de datos utilizados en las pruebas y los resultados obtenidos, se encuentra disponible en el repositorio de GitHub. Esto facilita su acceso a otros investigadores interesados en analizar, replicar o mejorar el sistema desarrollado.

Conclusiones

El presente trabajo confirmó que es posible desarrollar una herramienta eficiente para la gestión y análisis de documentos PDF en entornos con recursos limitados, como el cubano, utilizando la técnica Retrieval-Augmented Generation (RAG). El prototipo evaluado mostró un rendimiento satisfactorio en términos de precisión y rapidez, validando su viabilidad en escenarios de baja infraestructura tecnológica. Esta herramienta ofrece una alternativa local a soluciones comerciales inaccesibles, promoviendo la soberanía tecnológica al reducir la dependencia de plataformas extranjeras. El prototipo tiene un gran potencial para futuras aplicaciones, como en plataformas educativas nacionales, donde puede facilitar el acceso a recursos de estudio en contextos con limitación de profesores, y en otros sectores como el asesoramiento médico o la gestión de quejas empresariales. Entre las líneas futuras de desarrollo se propone la evaluación con conjuntos de datos más amplios y la incorporación de modelos más pequeños o multimodales que permitan ampliar sus capacidades, incluyendo la interpretación de imágenes y gráficos. Estos avances aumentan la versatilidad del sistema y extenderían su impacto en diversos sectores

Referencias

Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. RQ-RAG: Learning to Refine Queries for Retrieval Augmented Generation, 2024. URL https://arxiv.org/abs/2404.00610. Version Number: 1.

Andrew M. Dai, Christopher Olah, and Quoc V. Le. Document embedding with paragraph vectors, 2015. URL https://arxiv.org/abs/1507.07998.

Yorlis Delgado López. Digital documents and archival legislation: the case of mexico and cuba. *Investigación bibliotecológica*, 36(90), 2022.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library, 2024. URL https://arxiv.org/abs/2401.08281.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024. URL https://arxiv.org/abs/2312.10997.

Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. Chartllama: A multimodal llm for chart understanding and generation, 2023. URL https://arxiv.org/abs/2311.16483.

Yizheng Huang and Jimmy Huang. A survey on retrieval-augmented text generation for large language models, 2024. URL https://arxiv.org/abs/2404.10981.

Hoesung Lee, Katherine Calvin, Dipak Dasgupta, Gerhard Krinner, Aditi Mukherji, Peter Thorne, Christopher Trisos, José Romero, Paulina Aldunce, Ko Barrett, et al. *Climate change 2023: synthesis report. Contribution of working groups I, II and III to the sixth assessment report of the intergovernmental panel on climate change*. The Australian National University, 2023.

Research Meta. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models, September 2024. URL https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/.

Kurnia Muludi, Kaira Milani Fitria, Joko Triloka, and Sutedi. Retrieval-augmented generation approach: Document question answering using large language model. *International Journal of Advanced Computer Science and Applications*, 15(3), 2024. doi: 10.14569/IJACSA.2024.0150379. URL http://dx.doi.org/10.14569/IJACSA.2024.0150379.

Pouria Omrani, Alireza Hosseini, Kiana Hooshanfar, Zahra Ebrahimian, Ramin Toosi, and Mohammad Ali Akhaee. Hybrid retrieval-augmented generation approach for llms query response enhancement. In *2024 10th International Conference on Web Research (ICWR)*, pages 22–26. IEEE, 2024.

Giovanni Parodi and Cristóbal Julio. No solo existen palabras en los textos escritos: algunas teorías y modelos de comprensión de textos multimodales o multisemióticos. *Investigaciones sobre lectura*, 8:27–48, 2017.

Spurthi Setty, Harsh Thakkar, Alyssa Lee, Eden Chung, and Natan Vidra. Improving retrieval for rag based question answering models on financial documents, 2024. URL https://arxiv.org/abs/2404.07221.

Kaize Shi, Xueyao Sun, Qing Li, and Guandong Xu. Compressing long context for enhancing rag with amr-based concept distillation, 2024. URL https://arxiv.org/abs/2405.03085.

María Eugenia Toledo-Romaní, Mayra García-Carmenate, Carmen Valenzuela-Silva, Waldemar Baldoquín-Rodríguez, Marisel Martínez-Pérez, Meiby Rodríguez-González, Beatriz Paredes-Moreno, Ivis Mendoza-Hernández, Raúl González-Mujica Romero, Oscar Samón-Tabio, et al. Safety and efficacy of the two doses conjugated protein-based soberana-02 covid-19 vaccine and of a heterologous three-dose combination with soberana-plus: a double-blind, randomised, placebo-controlled phase 3 clinical trial. *The Lancet Regional Health–Americas*, 18, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Hanjie Zhang and Peter Kotanko. # 1506 uremic toxicity: gaining novel insights through ai-driven literature review. *Nephrology Dialysis Transplantation*, 39(Supplement_1):gfae069–0657, 2024.

Conflicto de interés

El autor autoriza la distribución y uso de su artículo.

Contribuciones de los autores

1. Conceptualización: Angel Alberto Vazquez Sánchez.
2. Curación de datos: Joaquin Rivas Sánchez.
3. Análisis formal: Naylin Brizuela Capote, Joaquin Rivas Sánchez.
4. Investigación: Joaquin Rivas Sánchez.
5. Metodología: Naylin Brizuela Capote.
6. Administración del proyecto: Angel Alberto Vazquez Sánchez.
7. Software: Joaquin Rivas Sánchez, Naylin Brizuela Capote.
8. Supervisión: Angel Alberto Vazquez Sánchez.

9. Validación: Naylin Brizuela Capote.

10. Visualización: Joaquin Rivas Sánchez.

11. Redacción - borrador original: Joaquin Rivas Sánchez, Naylin Brizuela Capote.

12. Redacción - revisión y edición: Naylin Brizuela Capote, Joaquin Rivas Sánchez.