

Diseño de una solución de búsqueda de información documental con modelos LLM y Embedding-Based Retrieval, para atención al cliente en la empresa Pública de Aseo y Gestión Ambiental del cantón Latacunga – EPAGAL.

Salgado Quimbita, Juan Mauricio

Vicerrectorado Académico General

Centro de Posgrados

Maestría en Ingeniería de Software

Tesis, previo a la obtención del título de Magister en Ingeniería de Software

Mgs. Gallardo Corrales, Diego Eduardo

25 de agosto del 2026



Daky Rea Vaca

SLAGADO JUAN MAURICIO - TESIS - revisado V9_ksi Antiplagio.pdf

Comprueba la originalidad de tu deber antes de enviarlo en tu curso - TURNITIN

202510 Turnitin-D2L-Respondus para el estudiante

Universidad San Francisco de Quito

Detalles del documento

Identificador de la entrega
trn:oid:::1:3372986997

Fecha de entrega
14 oct 2025, 10:35 a.m. GMT-5

Fecha de descarga
14 oct 2025, 10:46 a.m. GMT-5

Nombre del archivo
SLAGADO_JUAN_MAURICIO_-_TESIS_-_revisado_V9_ksi_Antiplagio.pdf

Tamaño del archivo
5.5 MB

102 páginas

17.415 palabras

100.564 caracteres

turnitin Página 1 de 110 - Portada Identificador de la entrega trn:oid:::1:3372986997

turnitin Página 2 de 110 - Descripción general de integridad Identificador de la entrega trn:oid:::1:3372986997

7% Similitud general

El total combinado de todas las coincidencias, excluidas las fuentes superpuestas, para ca...

Fuentes principales

6% Fuentes de Internet

1% Publicaciones

2% Trabajos entregados (trabajos del estudiante)

Diego Eduardo Gallardo Corrales

Time Stamping Security Data

Mgs. Gallardo Corrales, Diego Eduardo

Director

C.C.: 0503201717

Vicerrectorado Académico General

Centro de Posgrados

Certificación

Certifico que la Tesis: "**Diseño de una solución de búsqueda de información documental con modelos LLM y Embedding-Based Retrieval, para atención al cliente en la empresa Pública de Aseo y Gestión Ambiental del cantón Latacunga – EPAGAL.**" fue realizado por el señor **Salgado Quimbita, Juan Mauricio;** el mismo que cumple con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, además fue revisado y analizado en su totalidad por la herramienta de prevención y/o verificación de similitud de contenidos; razón por la cual me permito acreditar y autorizar para que se lo sustente públicamente.

Latacunga, 25 de agosto del 2026

Diego Eduardo Gallardo Corrales
Time Stamping Security Data

Mgs. Gallardo Corrales, Diego Eduardo

Director

C.C.: 0503201717

Vicerrectorado Académico General

Centro de Posgrados

Responsabilidad de Autoría

Yo, **Salgado Quimbita, Juan Mauricio** con cédula de ciudadanía N° 0502410525, declaro que el contenido, ideas y criterios de la Tesis: **"Diseño de una solución de búsqueda de información documental con modelos LLM y Embedding-Based Retrieval, para atención al cliente en la empresa Pública de Aseo y Gestión Ambiental del cantón Latacunga – EPAGAL"** es de mí autoría y responsabilidad, cumpliendo con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Latacunga, 25 agosto del 2026

Validar únicamente en FirmaEC.
Firmado electrónicamente por:
JUAN MAURICIO SALGADO QUIMBITA

Salgado Quimbita, Juan Mauricio

C.C.: 0502410525

Vicerrectorado Académico General

Centro de Posgrados

Autorización de Publicación

Yo, **Salgado Quimbita, Juan Mauricio** con cédula de ciudadanía N° 0502410525, autorizo a la Universidad de las Fuerzas Armadas ESPE, publicar la Tesis: ***"*Diseño de una solución de búsqueda de información documental con modelos LLM y Embedding-Based Retrieval, para atención al cliente en la empresa Pública de Aseo y Gestión Ambiental del cantón Latacunga – EPAGAL"** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mí/nuestra responsabilidad.

Latacunga, 25 de agosto del 2026

Validar únicamente en FirmaEC.
Firmado electrónicamente por:
JUAN MAURICIO
SALGADO QUIMBITA

Salgado Quimbita, Juan Mauricio

C.C.: 0502410525

Dedicatoria

Dedico este trabajo, ante todo, a mi madre, quien ha sido el pilar más firme en mi vida. Su amor incondicional, su fuerza silenciosa y su ejemplo de perseverancia me han guiado en cada paso de este camino. A ella le debo la inspiración para continuar, incluso cuando las circunstancias parecían difíciles. Su apoyo constante y sus palabras de aliento han sido el motor que me ha impulsado a alcanzar este logro que también es suyo.

A mis hermanas, quienes, con su cariño, comprensión y compañía, han llenado de alegría y equilibrio los momentos más intensos de este proceso. Gracias por creer en mí, por estar presentes siempre y por recordarme la importancia de la unión familiar y el esfuerzo compartido.

A todas las personas que, de una u otra forma, han sido parte de mi vida académica y profesional, especialmente a quienes compartieron este camino en la Empresa Pública de Aseo y Gestión Ambiental del Cantón Latacunga – EPAGAL. Su colaboración y experiencias fueron una fuente de aprendizaje constante que enriqueció este trabajo.

Finalmente, dedico este logro a todos quienes creen en el poder del conocimiento y el esfuerzo colectivo para transformar las instituciones públicas en espacios más eficientes, humanos y comprometidos con la comunidad.

Agradecimiento

Expreso mi más sincero agradecimiento a todas las personas e instituciones que hicieron posible la realización de este trabajo de investigación. En primer lugar, mi gratitud al Mgtr. Diego Gallardo, por su orientación constante, su paciencia y su valioso acompañamiento durante el desarrollo de este estudio. Sus observaciones y sugerencias fueron esenciales para fortalecer el rigor metodológico y enriquecer la calidad del documento.

A la Universidad de las Fuerzas Armadas ESPE Latacunga, por brindarme la oportunidad de ampliar mis conocimientos y desarrollar una visión crítica y propositiva sobre el tema. A mis docentes, quienes, con su compromiso y experiencia, contribuyeron significativamente a mi formación profesional y personal. Mi reconocimiento también a la Empresa Pública de Aseo y Gestión Ambiental del Cantón Latacunga – EPAGAL, por haberme permitido realizar esta investigación en su entorno institucional. Agradezco de manera especial a los directivos, jefaturas y compañeros de trabajo que colaboraron con información, tiempo y experiencias que enriquecieron los resultados de este estudio.

A mi familia, por su apoyo incondicional, su comprensión y su confianza en cada paso de este proceso. A mi madre y mis hermanas, por ser la fuerza que me impulsa y el ejemplo que me guía; y a mi pareja, por su compañía constante y su motivación en los momentos de mayor exigencia. Finalmente, agradezco a todas las personas que, de alguna manera, contribuyeron a que este proyecto se concrete. Este trabajo no solo representa un logro académico, sino también un compromiso con la innovación, la transparencia y la mejora continua en la gestión pública.

Índice de contenidos

Caratula 1
Reporte de verificación de contenidos 2
Certificación 3
Responsabilidad de Autoría 4
Autorización de Publicación 5
Dedicatoria 6
Agradecimiento 7
Índice de contenidos 8
Índice de tablas 13
Índice de figuras 14
Resumen 15
Abstract 16
Capítulo I 17
Generalidades 17
Introducción 17
Antecedentes 17
Planteamiento del problema 20
Objetivos 23
Objetivo General 23
Objetivos Específicos 23
Justificación e importancia 23
Alcance de la investigación 25
Hipótesis 26

Variables de la investigación 26

Variable Independiente 26

Conceptualización de la variable independiente 26

Variable Dependiente 27

Operacionalización de la variable dependientes (Indicadores) 27

Capítulo II 29

Marco Teórico 29

Gestión documental y Marco DAMA 29

Marco DAMA-DMBOK 29

Concepto de Gestión Documental *31*

Sistema de Gestión Documental 31

1960: Archivos físicos y catálogos manuales. 33

1960 – 1990: Bases de datos relacionales y recuperación por palabras clave. 33

1990 – 2010: Web, motores de búsqueda y aprendizaje automático. 34

2010-2020: Modelos de lenguaje y búsquedas semánticas. 34

2020: Chatbots y RAG. *34*

Fundamentos de la Inteligencia Artificial 35

Procesamiento de Lenguaje Natural 35

Chatbots y Asistentes Virtuales 35

Modelos Neuronales de Lenguaje 36

Grandes Modelos de Lenguaje (LLM) 36

Modelos Representativos 37

Análisis Comparativo 38

Retrevial – Augmented Generation (RAG) 40

Arquitectura 41

Concepto y funcionamiento de embeddings 42

Bases de datos vectoriales 44

Agentes Conversacionales 47

IA Generativa 47

Arquitecturas de diálogo 47

Asistentes Inteligentes 48

Human-in-the-loop (HITL) 48

Aplicaciones y uso de IA en la Administración Pública 49

Tecnologías para el sistema propuesto 50

Lenguaje de Programación de Backend y Frontend 50

Framework para Implementación de RAG 50

Capitulo III 54

Propuesta Metodológica y Diseño de Solución de Búsqueda de Información Documental 54

Integración de RAG y Gestión Documental 54

Caso práctico: Empresa Pública de Aseo y Gestión Ambiental del Cantón Latacunga – EPAGAL 57

Arquitectura Propuesta 57

Componentes de la arquitectura *58*

Diseño del modelo de búsqueda documental 62

Capitulo IV 67

Implementación de la Propuesta de Solución 67

Fase 1: Preprocesamiento y Vectorización Documental 67

Fase 2: Limpieza y Segmentación de Fragmentos. 68

Fase 3: Vectorización Semántica (LangChain) 68

Fase 4: Indexación en ChromaDB 70

Fase 5: Recuperación y Generación de Respuestas 70

Fase 6: Construcción de la Interfaz Conversacional 72

Gemini LLM 73

Flujo Conversacional 75

Lineamientos de Despliegue y Monitoreo 76

Pruebas de Fábrica y Aceptación de Usuarios 84

Entorno y Metodología de Validación 87

Infraestructura de Pruebas 87

Pruebas de Backend 88

Pruebas Unitarias 89

Pruebas de Integración 92

Pruebas de Carga (Lcust) 95

Interpretación de la Figura 15 96

Interpretación de la Figura 16 98

Capítulo V 99

Métricas de Evaluación Aplicadas 99

Verificación de Hipótesis 100

Planteamiento de la Hipótesis 100

Resultados de la Experimentación 100

Interpretación 101

Capítulo VI 103

Conclusiones y recomendaciones 103

Conclusiones.....103

Recomendaciones.....105

Referencias.....106

Apéndices.....112

Índice de tablas

Tabla 1 Objetivos de la gestión documental según DAMA-DMBOK.....29

Tabla 2 Beneficios de aplicar DAMA-DMBOK en gestión documental.....30

Tabla 3 Comparativa de modelos LLM.....38

Tabla 4 Comparativa de Likert.....40

Tabla 5 Cuadro comparativo de bases de datos vectoriales.....44

Tabla 6 Comparativa con escala Likert.....46

Tabla 7 Elección de ChromaDB como base vectorial.....59

Tabla 8 Proceso de orquestación LangChain + Python.....64

Tabla 9 Fases de la recuperación semántica.....71

Tabla 10 Integración en sistema RAG.....75

Tabla 11 Escenarios Pruebas de Integración.....84

Tabla 12 Resultados por dimensiones.....86

Tabla 13 Componentes y herramientas para pruebas.....87

Tabla 14 Pruebas de enpoints API REST.....89

Tabla 15 Pruebas Unitarias.....90

Tabla 16 Datos para aplicar las métricas de evaluación.....99

Tabla 17 Resultado de la experimentación.....100

Índice de figuras

Figura 1 Componentes principales de la arquitectura–Reader 41

Figura 2 Arquitectura de referencia RAG 41

Figura 3 Integración EBR – RAG.....44

Figura 4 Funcionamiento EBR) 52

Figura 5 Sistema Inteligente de Búsqueda y Recuperación Documental EPAGAL.....53

Figura 6 Integración entre gestión documental y RAG 56

Figura 7 Arquitectura Sistema Inteligente 57

Figura 8 Mapeo de capas del modelo 62

Figura 9 Vectorización semántica con.....69

Figura 10 Construcción de.....74

Figura 11 Lineamientos de Despliegue y Monitoreo.....77

Figura 12 Sistema Inteligente de Recuperación Documental.....79

Figura 13 Normativas institucionales y ordenanzas.....80

Figura 14 Reglamentos interno del personal 81

Figura 15 Código de trabajo 82

Figura 16 Uniformes y gestión operativa.....83

Figura 17 Verificación de salud del sistema.....92

Figura 18 Consulta general de saludo.....93

Figura 19 Consulta directa sobre cronograma operativo.....93

Figura 20 Entrada vacía.....94

Figura 21 Carga de base de conocimiento en ChromaDB 95

Figura 22 Estadística de solicitudes.....96

Figura 23 Distribución de tiempos de respuestas.....97

Resumen

Los sistemas de búsqueda documental son herramientas fundamentales para encontrar información precisa, permitiendo el acceso a la información y representa un reto significativo por la cantidad creciente de datos y la urgencia de ofrecer respuestas ágiles a la ciudadanía. En el caso de la Empresa Pública de Aseo y Gestión Ambiental del cantón Latacunga (EPAGAL), se evidencio obstáculos para acceder de manera ágil y precisa a las normativas, procesos y regulaciones internas, lo que impacta directamente en la atención al usuario y la gestión documental. Por ello, el estudio tiene como objetivo diseñar un sistema inteligente de búsqueda documental basado en Modelos de Lenguaje de Gran Escala (LLM) y técnicas de Recuperación Basada en Embeddings (EBR), para optimizar la gestión de documentos y mejorar la atención de usuarios. Para alcanzar lo propuesto, se realiza una revisión de literatura científica y técnica, identificando tecnologías y conceptos sobre los sistemas de recuperación semántica. En este contexto, se diseñó una solución inteligente fundamentado en una arquitectura tipo RAG (Retrieval-Augmented Generation), integrando herramientas como LangChain, ChromaDB, FastAPI, sentence-transformers y el modelo Gemini. Estas procesan documentos en múltiples formatos, vectoriza y almacenan en bases de datos semánticas. Además, desarrolló un chatbot como interfaz de acceso para los usuarios. El resultado fue un sistema inteligente con la habilidad de responder preguntas en lenguaje natural mediante la recuperación de fragmentos documentales, proporcionando respuestas contextualizadas y comprensibles al usuario. Demostrando estabilidad operativa y coherencia en las respuestas de acuerdo a la base de conocimiento, validándose a través de métricas de precisión, recall y pruebas realizadas en escenarios reales con usuarios de EPAGAL. Para concluir, el sistema propuesto es una alternativa tecnológica para búsqueda de información y gestión documental, orientado a mejorar el servicio a los usuarios internos como a clientes de entidades públicas.

***Palabras clave**:* búsqueda de información, LLM, EBR, inteligencia artificial, chatbot, agente conversacional.

Abstract

Document retrieval systems are fundamental tools for finding accurate information, allowing access to information and representing a significant challenge due to the growing amount of data and the urgency of providing agile responses to citizens. In the case of the Public Company for Waste Management and Environmental Management of the Latacunga canton (EPAGAL), obstacles were evidenced in accessing regulations, processes, and internal regulations in an agile and accurate manner, which directly impacts user service and document management. Therefore, the study aims to design an intelligent document retrieval system based on Large Language Models (LLM) and Embedding-Based Retrieval (EBR) techniques, to optimize document management and improve user service. To achieve the proposed goal, a review of scientific and technical literature is carried out, identifying technologies and concepts on semantic retrieval systems. In this context, an intelligent solution was designed based on a Retrieval-Augmented Generation (RAG) architecture, integrating tools such as LangChain, ChromaDB, FastAPI, sentence-transformers, and the Gemini model. These process documents in multiple formats, vectorize and store them in semantic databases. In addition, a chatbot was developed as a user access interface. The result was an intelligent system with the ability to answer questions in natural language through the retrieval of document fragments, providing contextualized and understandable responses to the user. It demonstrated operational stability and coherence in the answers according to the knowledge base, being validated through precision and recall metrics and tests carried out in real scenarios with EPAGAL users. In conclusion, the proposed system is a technological alternative for information retrieval and document management, aimed at improving the service provided to internal users as well as clients of public entities.

Keywords: information search, LLM, EBR, artificial intelligence, chatbot, conversational agent.

Capítulo I

Generalidades

Introducción

En este capítulo se describen las generalidades, antecedentes históricos y evolución de los sistemas de búsqueda y recuperación de información. Con base a lo expuesto de los antecedentes históricos se procede a definir el planteamiento del problema que reflejan desde los métodos manuales tradicionales hasta las soluciones actuales basadas Modelos informáticos de amplio alcance diseñados para generar texto utilizando inteligencia artificial conocidos como LLM, así también corresponde al detalle de los objetivos para lograr el proyecto de investigación. Por lo tanto, es importante presentar la razón y la importancia del tema de investigación sugerido. Al final, se establecen las hipótesis, que se comprenden como afirmaciones que no han sido confirmadas y que, una vez que sean desmentidas o validadas, dejan de ser hipótesis y se transforman en afirmaciones verificadas. Bajo este contexto, el recorrido evolutivo justifica el diseño de una solución de búsqueda y recuperación de información sostenida en LLMs y EBR, constituyéndose una propuesta de innovación en el acceso de información.

Antecedentes

El crecimiento de la Inteligencia Artificial (IA) ha impregnado cada faceta de nuestra vida, moviendo a la humanidad hacia la cuarta revolución industrial (Gil Peñaranda, 2023) . La evolución de la Inteligencia Artificial Generativa (IAG) han potenciado el desarrollo de Grandes Modelos de Lenguaje (LLM) que muestran comportamiento emergente para los que fueron desarrollados. Desde que ChatGPT fue presentado en noviembre de 2022, los amplios modelos de lenguaje han mostrado un notable potencial en múltiples usos, abarcando desde la escritura de contenido y la creación de conceptos hasta la codificación. (Díaz de Rábago Pemán, 2024).

El Procesamiento de Lenguaje Natural se manifiesta en un gran número de sistemas informáticos que utilizamos día a día. Por lo antes mencionado, empleamos buscadores para recuperar y extraer información entre la cantidad masiva de contenido documental y textual, sistemas de traducción automática, corrección ortográfica y gramatical, agentes conversacionales para obtener información de servicios por voz o texto y sistemas de recomendación de contenidos (revisiones, noticias, páginas web, artículos) (Lara,2019).

Los Modelos de Lenguaje a Gran Escala (LLM) han revolucionado múltiples áreas, desde el procesamiento del lenguaje natural hasta la interacción conversacional con grandes volúmenes de datos. Estos modelos de inteligencia artificial creadas para entender, manejar y producir texto en un idioma humano. Son innovaciones significativas en el campo de la inteligencia artificial, debido a su capacidad para interactuar de manera concordante y fluida en lenguaje natural. Los LLM han alcanzado niveles de sofisticación que les permiten no solo responder preguntas o redactar texto, sino también interpretar matices y contextos complejos, lo que los convierte en herramientas versátiles y de amplio alcance (Martin de Pablo, 2024). Las aplicaciones de los LLM son numerosas y se extienden por diversos campos, desde la educación hasta la medicina, pasando por servicios de atención al cliente, análisis de datos o la creación de contenido. Los LLM representan un avance tecnológico fundamental en el procesamiento de lenguaje natural, impulsado por décadas de investigación y desarrollo en inteligencia artificial.

En este contexto, los LLM, como ChatGPT, acorde a los avances en inteligencia artificial y en el procesamiento del lenguaje natural, han mostrado un desarrollo notable recientemente. Estos modelos se aplican ampliamente en el desarrollo de chatbots y asistentes virtuales cuya finalidad es generar conversaciones naturales con los usuarios. Dichos sistemas se utilizan en servicios de atención al cliente, soporte técnico y consulta de información general (Henry & Figueroa, 2024).

Existen modelos que han sido entrenados con millones de parámetros para responder a la generación de texto en sus diferentes áreas. Sin embargo, muchos de ellos, una vez que han sido entrenados, no pueden acceder a nueva información, convirtiéndose

en una figura de información estática sin posibilidad de actualización. La imposibilidad de aceptar nueva información puede crear consultas erróneas, ya sea porque la documentación es antigua y se ha actualizado más tarde que el entrenamiento del modelo, o por la malinterpretación de datos. El uso de técnicas como RAG (Retrieval Augmented Generation) y EBR (Embedding Base Retrival) busca resolver estos problemas, entrenando modelo y ayudando a recuperar información (Elizalde Roldán, 2024). La técnica RAG integra la recuperación de información con la ampliación del contexto y generación de cadena más coherentes. Así también, La recuperación basada en incrustación a gran escala (EBR) constituye el fundamento esencial de las aplicaciones industriales vinculadas a la búsqueda.

En la actualidad, existen empresas que demanda de atención al cliente y sus respuestas sean claras y precisas. Por ello se ha decidido, enfatizar en la búsqueda de tecnologías como RAG el sistema de IA generativa que puedan ofrecer respuestas más precisas y contextualmente adecuadas. RAG ofrece una gran cantidad de aplicaciones prácticas en entornos empresariales particularmente en las áreas de la búsqueda de significado, la obtención de datos, la atención al cliente y la generación de contenido. Su capacidad para acceder a una amplia gama de datos y utilizarlos dinámicamente la convierte en una herramienta inestimable para las empresas que buscan optimizar diversas operaciones (Laparra Osca, 2024). Incorporar la tecnología EBR es un sofisticado mecanismo de IA que mejora la funcionalidad de los LLM mediante la integración de un sistema de recuperación dinámico. Este sistema permite a los LLM acceder a fuentes de datos externas y actualizadas y utilizarlas, enriqueciendo así sus respuestas con un mayor volumen de información.

La Empresa Pública de Aseo y Gestión Ambiental del cantón Latacunga (EPAGAL) realiza la gestión integral de los desechos sólidos generados dentro de su jurisdicción. En años recientes, ha quedado claro un aumento en las responsabilidades relacionadas con la administración de limpieza, al igual que un crecimiento en la cantidad de personas que usan el servicio. Sin embargo, las tecnologías implementadas para la atención ciudadana y la

búsqueda de información interna han quedado obsoletas, manteniéndose procesos manuales como la asignación de turnos o la solicitud de tickets para realizar trámites y consultas.

Por lo tanto, si se desea mejorar la atención al cliente, resulta necesario incorporar tecnologías de última generación, como los asistentes virtuales fundamentados en modelos lingüísticos que tienen la capacidad de contestar a las inquietudes de los usuarios en múltiples situaciones utilizando información de diferentes fuentes confiables (Cortés-Cediel et al., 2023) (Bettayeb et al., 2024) para incorporar enfoques como RAG (Retrieval-Augmented Generation) o EBR (Embedding-Based Retrieval) introduce un elemento de recuperación de información que emplea las solicitudes del usuario para obtener datos de fuentes recientes, ofreciendo de esta manera respuestas más actualizadas y precisas (Lewis et al., 2021;) (Gao et al., 2024) (Nie et al., 2025). Este enfoque, la pregunta del usuario y los datos pertinentes se entregan al modelo de lenguaje, el cual combina el nuevo conocimiento con sus datos de entrenamiento para generar respuestas más completas y contextualizadas.

De acuerdo a la evolución tecnológica, EPAGAL busca implementar un Bot Inteligente que pueda responder de manera eficiente a preguntas frecuentes de los usuarios y personal de la institución. Para ello, se busca crear un sistema basado en Modelos de Lenguaje de Gran Escala y Embedding-Based Retrieval. Así también, se pretende aplicar técnicas avanzadas de búsqueda semántica para garantizar respuestas precisas de los documentos internos relacionados con la gestión ambiental y residuos del cantón Latacunga.

Planteamiento del problema

En la actualidad, la gestión documental se ha logrado mantener debido a la necesidad de garantizar a los ciudadanos la información pública y estatal de calidad y optima (Jiménez Mocetón, 2022) . En el Ecuador, Los recientes enfoques en la

administración pública apuntan hacia la eficiencia, la efectividad, la claridad y la inclusión de la ciudadanía. Estas propuestas consideran las mejores experiencias de la gestión privada y las adaptan a su contexto, con el objetivo de ofrecer servicios públicos que respondan a las demandas de la comunidad utilizando avances tecnológicos. (Barragán Martínez, 2022).

Los sistemas tradicionales de búsqueda y consulta de información se han apoyado en bases de datos estructuradas y consultas manuales. Estas bases de datos organizan la información la información en filas y columnas, permitiendo la recuperación de datos por mediante lenguajes de consulta estructurada (Martínez López & Gallegos Ruiz, 2017). No obstante, este enfoque presenta limitaciones significativas: la precisión de la recuperación depende de la correcta formulación de las consultas, lo que con frecuencia conduce a resultados incompletos o irrelevantes, además de pérdidas de datos y tiempos de respuesta prolongados, especialmente al manejar grandes volúmenes de información.

En contraste, los enfoques actuales de recuperación de información se orientan hacia datos no estructurados como documentos textuales, reportes o normativas que requieren técnicas más avanzadas. Por ello, ha producido una evolución hacia sistemas de búsqueda basadas en inteligencia artificial. En este escenario, los Modelos de Lenguaje de Gran Escala y las técnicas de recuperación fundamentadas en embeddings han cambiado radicalmente la forma en que se maneja y se obtiene información (Valdés Parra, 2024). Los sistemas antes mencionados tienen la capacidad de entender el contexto y el significado semántico de las consultas, lo cual ayuda a una recuperación de información precisa y pertinente (Negueruela Gómez, 2024). A diferencia de las consultas estructuradas, estas tecnologías permiten comprender el contexto y el significado semántico del lenguaje natural, optimizando la precisión y pertinencia de las respuestas (Negueruela Gómez, 2024).

De este modo, la adopción de LLM y *embeddings* posibilita la recuperación en tiempo real de información documental a gran escala, disminuyendo los tiempos de búsqueda y mejorando la eficiencia en la gestión de datos, lo que constituye un avance clave frente a las limitaciones de los sistemas tradicionales.

Por ello, la institución pública de Aseo y Gestión Ambiental del Cantón Latacunga (EPAGAL), debido a la orientación empresarial, maneja grandes volúmenes de información documental relacionados con normativas, procedimientos, reglamentos y procesos de atención al cliente. Sin embargo, la gestión de la documentación enfrenta retos significativos en el consumo, consulta y revisión de información, lo que afecta la eficiencia en la atención al usuario. Uno de los problemas relevantes es el acceso ágil a la documentación oficial e interna, debido a que todavía se depende mucho de procesos manuales en la administración de la información. Esto ha provocado respuestas inexactas o demoradas a las preguntas, afectando de manera adversa la calidad del servicio.

Es así, que evidencia la necesidad de avanzar hacia la democratización de la información, entendida como la capacidad de exponer de manera eficiente el conocimiento contenido en los documentos institucionales y ponerlo a disposición de todos los usuarios de forma rápida y precisa. Para enfrentar estos desafíos, resulta necesario implementar una solución tecnológica documental basada en Inteligencia Artificial, apoyada en Modelos de Lenguaje de Gran Escala (LLM) y en técnicas de Embedding-Based Retrieval (EBR), que optimicen el acceso, la recuperación y el aprovechamiento de la información institucional.

El cantón Latacunga, cuenta con una Empresa Pública de Aseo y Gestión Ambiental (EPAGAL), fundada el 20 de julio del 2010, la cual constituye una entidad clave en la gestión de residuos y en la gestión ambiental. Debido a la creciente concurrencia de usuarios y al aumento de información, la institución enfrenta desafíos importantes en la gestión de la información documental y en la atención al cliente por la ausencia de sistemas avanzados de búsqueda. La infraestructura tecnológica actual para estos procesos presenta limitaciones, ya que la entidad no dispone de un sistema de gestión documental centralizado y digitalizado, lo que obliga a que gran parte de las actividades se realicen de manera manual.

Bajo este contexto, la implementación de sistemas de búsqueda basados en inteligencia artificial ayudará a mejorar la recuperación de información, incrementando la eficiencia en la atención a los usuarios y empleados. Hoy en día, EPAGAL no dispone de un

sistema de este tipo, lo cual limita su capacidad de respuesta ante las consultas. Incorporar herramientas avanzadas de búsqueda documental permitiría el almacenamiento centralizado de documentos, y facilitaría el acceso a la información. La carencia de este tipo de sistemas ha limitado la eficiencia institucional y calidad del servicio ofrecido al ciudadano.

En base a lo expuesto, se plantea el siguiente problema:

¿Cómo mejorar la atención al cliente en EPAGAL mediante el uso de Inteligencia Artificial, optimizando la búsqueda y recuperación de información en las normativas, reglamentos y operaciones de la institución?

Objetivos

Objetivo General

Diseñar un sistema inteligente con interfaz conversacional de búsqueda y recuperación de información documental basado en Modelos de Lenguaje de Gran Escala (LLM) y técnicas de Embedding-Based Retrieval (EBR) para mejorar la gestión documental interna y la atención al cliente.

Objetivos Específicos

Diseñar una solución tecnológica búsqueda documental para integrar técnicas de recuperación basada en embeddings (EBR) y Modelos LLM.

Desarrollar un sistema inteligente basado en un Agente Conversacional, que implemente la solución de recuperación documental permitiendo la interacción mediante lenguaje natural.

Evaluar el desempeño del sistema inteligente para conocer los tiempos de respuesta en la búsqueda de información y satisfacción del usuario.

Justificación e importancia

En la actualidad, los LLM Modelos de Lenguaje a Gran Escala han transformado diversas disciplinas, abarcando desde el procesamiento del lenguaje natural hasta la interacción conversacional, todo ello manejando grandes volúmenes de datos. Por esta

razón, resulta fundamental modernizar el acceso a la información mediante el uso de Modelos de Lenguaje de Gran Escala y técnicas de Recuperación Basada en Embeddings. Estas tecnologías no solo permiten interpretar el lenguaje natural, sino que también optimizan la precisión en la recuperación de documentos.

El acceso a la información en EPAGAL, Se ha vuelto una exigencia esencial, ya que la efectividad en la obtención de documentos afecta de manera directa la calidad del servicio brindado a los usuarios y la localización de información por parte del personal administrativo. Sin embargo, la empresa maneja grandes volúmenes de información documental que no son accesibles de forma tradicional, lo cual ocasiona retrasos en la atención al usuario y una enorme carga administrativa.

Esto genera diversos problemas, tales como retrasos en la obtención de documentos oficiales que afectan los tiempos de respuesta; inconsistencias en la información proporcionada, lo que genera incertidumbre entre la ciudadanía; y dificultades en la búsqueda de normativas y regulaciones, lo que afecta de manera desfavorable la elección de decisiones en lo que respecta a la administración del medio ambiente y la limpieza pública.

En este contexto, la implementación de Modelos de Lenguaje de Gran Escala (LLM) y de sistemas de recuperación basados en embeddings se presenta como una solución innovadora. Estas tecnologías permiten interpretar consultas en lenguaje natural, ofreciendo una experiencia de búsqueda intuitiva y efectiva. Así mismo, la recuperación de información garantiza resultados precisos en menor tiempo, optimizando los procesos administrativos y asegurando un servicio eficiente y accesible.

La presente investigación servirá como documento base en el desarrollo e implementación de tecnologías avanzadas en la recuperación de información en instituciones públicas, con la finalidad de mejorar el acceso y gestión de datos. Por tal razón, los beneficiaros directos de la investigación incluye al personal de EPAGAL encargados de la gestión de información documental en el ámbito de normativas, regulaciones y procesos. Así también, los funcionarios encargados de la atención al cliente,

utilizaran la solución para responder las consultas de los usuarios de forma rápida y precisas; cabe destacar que la investigación servirá a especialistas en inteligencia artificial y recuperación de información, que se encuentren interesados en aplicar Modelos de Lenguaje de Gran Escala (LLM) y Embedding-Based Retrieval en sistemas documentales. Así también, se beneficiarán las instituciones públicas que están en investigaciones de mejorar sus sistemas de búsqueda documental podrán adoptar la solución propuesta. En conclusión, este estudio no solo influirá en la implementación tecnológica inmediata, lo cual con lleva a un punto de partida para futuras investigaciones sobre la búsqueda semántica en entidades gubernamentales.

La implementación de esta propuesta conllevará una serie de beneficios. Para iniciar, se realizará la recuperación de información con la integración de técnicas de inteligencia artificial, lo que permitirá un acceso ágil y controlado a documentos normativos y administrativos dentro de EPAGAL. En este sentido, se automatizará el proceso de búsqueda documental, eliminando la dependencia de métodos manuales y reduciendo significativamente los tiempos de respuesta.

Asimismo, la solución incorporará mecanismos de seguridad y perfiles de acceso diferenciados, garantizando que cada usuario visualice únicamente la información correspondiente a sus funciones. Esto no solo mejorará la atención al cliente al proporcionar respuestas más rápidas y precisas, sino que también apoyará al personal interno en la estandarización de procesos y consultas documentales confiables.

Finalmente, la propuesta tiene un impacto escalable y puede ser replicada en otras instituciones públicas, facilitando su implementación en organismos gubernamentales con necesidades similares. De esta manera, el uso de tecnologías avanzadas contribuirá a la modernización y a la gestión segura de la información documental en el sector público.

Alcance de la investigación

La investigación se centra en el diseño, desarrollo e implementación de una interfaz conversacional con búsqueda y recuperación documental basado en Modelos de Lenguaje

de Gran Escala (LLM) y técnicas de Embedding-Based Retrieval (EBR), que será implementado en el contexto de EPAGAL. La solución abarcará la creación de una base de conocimientos sobre gestión de normativas, reglamentos y procesos administrativos cuyo objetivo es ayudar en la atención al cliente y apoyar a funcionarios internos en la consulta de información oficial. El sistema contará con una interfaz que permite la interacción conversacional para realizar consultas en lenguaje natural, además de métricas que evalúan la precisión en la recuperación, los tiempos de respuesta y la satisfacción de los usuarios. También incluirá dos perfiles de acceso para enmarcar la seguridad y el control. No obstante, en esta etapa no se incluye la integración total con todos los sistemas externos de la institución ni la automatización de los procesos administrativos; se enfoca en mejorar la recuperación y disponibilidad de la información documental. Por último, la propuesta será susceptible de ser replicada y ampliada en otras entidades públicas que posean necesidades similares, estableciendo un fundamento para futuras investigaciones en búsqueda semántica y asistencia conversacional en el ámbito gubernamental.

Hipótesis

Si se diseña una solución de búsqueda de información documental con modelos de Lenguaje de Gran Escala (LLM) y técnicas de *Embedding-Based Retrieval (EBR)*, entonces se optimizará la atención al cliente y la gestión documental en la Empresa Pública de Aseo y Gestión Ambiental del cantón Latacunga.

Variables de la investigación

Variable Independiente

Implementación de una solución de búsqueda de información documental basada en modelos LLM y Embedding Based Retrival.

Conceptualización de la variable independiente

Una solución de búsqueda de información documental basada en Modelos de Lenguaje de Gran Escala (LLM) y en técnicas de Embedding-Based Retrieval (EBR) integra

la inteligencia artificial, modelos de recuperación semántica y una interfaz conversacional. Su objetivo es optimizar la búsqueda y recuperación de información, utilizando agentes conversacionales para ofrecer respuestas precisas y agiles al usuario interno y clientes.

Variable Dependiente

Optimización de la atención al cliente en la Empresa Pública de Aseo y Gestión Ambiental del cantón Latacunga por medio de respuestas rápidas, contextualizadas y en lenguaje natural.

Operacionalización de la variable dependientes (Indicadores)

1. Tiempo promedio de respuesta del sistema a las consultas de usuarios (en segundos).
2. Precisión de las respuestas entregadas por el sistema (porcentaje de aciertos respecto a la base documental).
3. Nivel de satisfacción de los usuarios internos y externos (medido con encuestas tipo SUS o escala Likert).
4. Reducción de procesos manuales en la búsqueda de información documental (comparación antes/después de la implementación).
5. Accesibilidad a documentos normativos y administrativos (porcentaje de consultas resueltas correctamente).
6. Estabilidad y disponibilidad del sistema (porcentaje de tiempo en línea sin fallos).

Para la operalización de la variable dependiente enfocado a la optimización de la atención al cliente y gestión documental integrado en un sistema inteligente, se sustentará mediante las siguientes métricas:

Precisión (Precisión)

$$Precision = \frac{TP}{TP+FP}$$

TP (True Positives): Documentos relevantes recuperados.

FP (False Positives): Documentos irrelevantes recuperados.

La precisión refleja el nivel de exactitud de los resultados que se logran, lo cual significa cuán fiable es el sistema al proporcionar datos pertinentes.

Recuperación (Recall)

$$Recall = \frac{TP}{TP + FN}$$

FN (False Negatives): Documentos relevantes que no fueron recuperados.

El recall evalúa la habilidad del sistema para localizar la mayor cantidad de documentos pertinentes, analizando qué proporción de la información significativa fue realmente reconocida.

F1-Score (Promedio armónico de Precisión y Recuperación)

$$F1 = 2\,x\,\frac{Precision\ x\ Recall}{Precision + Recall}$$

El F1-Score se utiliza para lograr un equilibrio entre la precisión y la recuperación, especialmente en escenarios donde ambos indicadores son importantes y deben considerarse conjuntamente.

Capitulo II

Marco Teórico

Gestión documental y Marco DAMA

Marco DAMA-DMBOK

Bajo este contexto DAMA-DMBOK (Cuerpo de Conocimiento en Gestión de Datos), dado a conocer por la Asociación Internacional de Gestión de Datos (DAMA), se establece como una referencia ampliamente utilizada para la administración de datos y documentos. Este modelo sugiere un enfoque holístico, donde la gestión de documentos se conecta con otras funciones de gobernanza de datos como la estructura, la calidad, los metadatos y la protección (Mosley et al., 2010).

A continuación, se observa la Tabla 1, sobre los objetivos de la gestión documental según el marco DAMA-DMBOK tienen como objetivo asegurar que los datos y documentos estén disponibles, sean de calidad, seguros y se puedan seguir a lo largo de su proceso, garantizando que se utilicen de manera fiable y en el momento adecuado dentro de la entidad.

Tabla 1

Objetivos de la gestión documental según DAMA-DMBOK

Objetivo	Descripción
Disponibilidad	Garantizar que los documentos estén accesibles a los usuarios autorizados en el momento y formato adecuados.
Calidad	Mantener la precisión, coherencia y actualidad de los documentos, evitando duplicidades o versiones contradictorias.

Objetivo	Descripción
Seguridad	Asegurar que la información confidencial esté a salvo de accesos no permitidos, extravíos o modificaciones inapropiadas.
Trazabilidad	Documentar el desarrollo de cada archivo, abarcando desde su generación hasta su eliminación definitiva.
Eficiencia	Optimizar la recuperación de documentos mediante herramientas tecnológicas, clasificadores y metadatos.

Nota. Tomado de Mosley et al. (2010).

En la Tabla 2, se pueden notar las ventajas del marco DAMA-DMBOK en la gestión de documentos, ya que facilita una mejor toma de decisiones al contar con información veraz y reciente.

Tabla 2

Beneficios de aplicar DAMA-DMBOK en gestión documental

Beneficio	Descripción
Mejor toma de decisiones.	Permite contar con información confiable, precisa y actualizada, reduciendo la incertidumbre en los procesos.
Agilidad en la atención ciudadana.	Disminuye los tiempos de búsqueda de normativa, reglamentos o procesos, mejorando la eficiencia en la respuesta.
Transparencia y rendición de cuentas.	Facilita el acceso estructurado a los documentos oficiales, incrementando la confianza en la gestión pública.
Soporte a la transformación digital.	Integra la gestión documental con tecnologías avanzadas como IA, bases vectoriales y recuperación semántica.

Nota. Tomado de Mosley et al. (2010).

Concepto de Gestión Documental

La gestión documental es un área fundamental en la gobernanza de datos, ya que garantiza que la información que una entidad genera, recibe y utiliza esté correctamente estructurada, segura, accesible y disponible a lo largo de su ciclo vital (International Organization for Standardization [ISO], 2016).

Sistema de Gestión Documental

En la entidad pública de limpieza y manejo ambiental de Latacunga (EPAGAL), la organización de documentos no está regulada dentro de un esquema formal como DAMA-DMBOK. Los procedimientos en uso dependen del almacenamiento manual en archivos tanto físicos como digitales, careciendo de una clasificación coherente y de un sistema centralizado para su recuperación.

Las principales limitaciones identificadas son:

1. Recuperación lenta de documentos normativos y operativos.

2. Dependencia del conocimiento del personal, lo cual dificulta el acceso en ausencia de actores clave.

3. Falta de trazabilidad y control de versiones, generando riesgos de uso de información desactualizada.

Este diagnóstico evidencia la necesidad de evolucionar hacia un sistema de gestión documental moderno, donde la información pueda ser consultada en lenguaje natural, organizada en bases vectoriales y optimizada con técnicas de inteligencia artificial. La adopción de un modelo DAMA-DMBOK permitiría formalizar procesos de gobierno documental y servir de base para la mejora que se propone con el chatbot inteligente y la integración de RAG (Retrieval-Augmented Generation).

La transformación de los sistemas de búsqueda de información documental está marcada por los desarrollos tecnológicos que han cambiado la manera en que las organizaciones almacenan, obtienen y gestionan información. Al principio, la búsqueda de documentos se basaba en técnicas manuales, tales como archivos en papel y catálogos indexados, lo cual limitaba el acceso inmediato a los datos. Con la llegada de las bases de datos estructuradas en la segunda parte del siglo XX (Piñero Gómez, 2014), se pusieron en marcha sistemas fundamentados en SQL y recuperación basada en palabras clave, incrementando la eficacia, pero con restricciones en la comprensión del lenguaje natural.

Desde los años 90, la recuperación de datos y procesamiento de lenguaje natural (PLN) (Periñán Pascual & Arcas Túnez, 2011), ayudo a la creación de motores de búsqueda avanzados. Un ejemplo de lo mencionado es Google, se implementó PageRank conocida como “Google Page Rank”, cuya funcionalidad era segmentar documentos en la web. Sin embargo, estos sistemas se basan en palabras clave precisas, siendo complejo la recuperación exacta de datos en documentos amplios con grandes volúmenes de datos para la época (Jara Moya et al., 2023).

Es así, para iniciar de la década del 2010, el avance de la Inteligencia Artificial y la Inteligencia Artificial Generativa estableció un hito importante gracias a la expansión y accesibilidad de datos, transformando el uso de redes neuronales y técnicas de aprendizaje profundo. Más específicamente, los LLM (Modelos de Lenguaje Grande) son sistemas de inteligencia artificial que han sido alimentados con amplias cantidades de texto. Tienen la habilidad de generar textos que suenan sorprendentemente humanos, lo que se debe a su capacidad de aprender a partir de patrones del lenguaje, información general y ciertos tipos de razonamientos a partir del material que se utilizó en su entrenamiento.

Por último, en el año 2022 han desarrollado herramientas informáticas que poseen algún tipo de características inteligentes. Esto se debe a la utilización de Modelos de Lenguaje de Gran Escala (LLM), en donde el concepto de “grande” hace referencia al

crecimiento de la inteligencia artificial a medida que aumenta la cantidad de datos y la capacidad de entrenamiento con extensas bases de conocimiento.

Uno de los modelos más representativos es GPT (Generative Pre-trained Transformer) (Brown et al., 2020), que ha alcanzado notoriedad por su capacidad para generar respuestas coherentes y naturales. Sin embargo, el GPT es solo uno de los múltiples LLM existentes, todos ellos basados en la arquitectura Transformer (García-Peñalvo et al., 2023). De manera complementaria, la Recuperación Basada en Embeddings (Embedding-Based Retrieval, EBR) constituye una técnica que potencia la búsqueda documental al establecer coincidencias semánticas entre consultas y documentos. Estas tecnologías están siendo aplicadas en distintos contextos como la sociedad, la salud, la educación y la gestión pública, entre otros.

Cabe mencionar que integrar Modelos de Lenguaje de Gran Escala (LLM) y técnicas de Recuperación Basada en Embeddings (EBR) en un sistema de búsqueda documental, junto con un chatbot inteligente ayudará, permitirá optimizar la recuperación de la información y atención al cliente.

Por lo tanto, es necesario mostrar la evolución de gestión documental para mostrar la evolución de los sistemas de búsqueda y gestión documental:

1960: Archivos físicos y catálogos manuales.

La gestión documental se basó en archivos físicos y catálogos manuales, tenía acceso restringido y una recuperación lenta e ineficiente de la información (Montilla Peña & Mena Mujica, 2013).

1960 – 1990: Bases de datos relacionales y recuperación por palabras clave.

Se expone el concepto de bases de datos relacionales y el lenguaje SQL, que facilita la automatización de trámites documentales. Surgen los primeros sistemas de recuperación

de información y se basan en palabras claves e indexación, sin capacidad de interpretar sinónimos o ambigüedades (Meijer & Bierman, 2011). (Chowdhury, 2010).

1990 – 2010: Web, motores de búsqueda y aprendizaje automático.

Con el auge de la web surgen motores de búsqueda Google que incorpora el algoritmo PageRank, esta transforma la clasificación de documentos acordes a la relevancia (Brin & Page, 1998; Olivas Varela, 2011). En paralelo, el procesamiento de lenguaje natural (PNL) se fortalece y se incorporan modelos de aprendizaje automático como k-NN, SVM y Naive Bayes, aplicados a clasificación y filtrado de documentos (Jurafsky et al., 2000).

2010-2020: Modelos de lenguaje y búsquedas semánticas.

Los Modelos de Lenguaje de Gran Escala (LLMs) basados en arquitecturas tipo Transformer (Vaswani et al., 2017), representan un cambio radical. Destaca BERT (Devlin et al., 2019), que presenta un entrenamiento en dos direcciones para captar el contexto semántico. Los modelos generativos como GPT-2 y GPT-3 (Brown et al., 2020) amplían las habilidades para interactuar en conversaciones. Al mismo tiempo, las técnicas de Recuperación Basada en Embeddings (EBR) posibilitan búsquedas semánticas a través de representaciones vectoriales, lo que facilita la identificación de coincidencias conceptuales incluso sin que las palabras coincidan de manera literal (Huang et al., 2020).

2020: Chatbots y RAG.

Con el lanzamiento de ChatGPT (2022), surge una nueva era de búsqueda conversacional en lenguaje natural. La integración de LLMs con EBR y arquitecturas híbridas como RAG (Retrieval-Augmented Generation) se ha aplicado en bibliotecas digitales, instituciones públicas, educación y entornos legales, permitiendo un acceso más contextualizado y eficiente a la información (Lima, 2024; Gao et al., 2024).

Fundamentos de la Inteligencia Artificial

Procesamiento de Lenguaje Natural

El Procesamiento del Lenguaje Natural es una parte de la inteligencia artificial y una sección de la lingüística informática que utiliza una variedad de procedimientos de programación estructurada para facilitar la comunicación con las personas en su idioma natal (García Reina, 2018).

En el estudio conceptual de la PNL, se trata de un área tecnológica que intenta conectar la inteligencia artificial (IA) con el lenguaje que empleamos en nuestra vida diaria. A través del aprendizaje automático y diversas áreas de la IA, mediante el uso de algoritmos, podemos entender los procesos que seguimos los humanos al comunicarnos. Generalmente, estos algoritmos se aplican en la traducción de documentos, la elaboración de resúmenes, el reconocimiento de voz, entre otros.

Chatbots y Asistentes Virtuales

Los chatbots son aplicaciones informáticas que, mediante el uso de aprendizaje automático y análisis de lenguaje natural, logran replicar una variedad de respuestas lógicas y, en un contexto específico, llevan a cabo una conversación similar a la humana (García Reina, 2018).

Según Ramírez Hernández & Valle Cruz (2022), los asistentes virtuales son agentes conversacionales creados con inteligencia artificial que utilizan el procesamiento del lenguaje natural para comunicarse con los usuarios y atender sus preguntas.

Actualmente, entidades gubernamentales han desarrollado chatbots que son de utilidad en diferentes áreas, principalmente en la atención al cliente. Un asistente virtual se define como un sistema que ha progresado paralelamente a los desarrollos en inteligencia artificial, aprendizaje automático y análisis de lenguaje humano. El desarrollo, es acelerado ya que es posible categorizarlos según el servicio que prestan. Las categorías comunes se

encuentran los chatbots de ventas, los chatbots de atención al cliente y los chatbots de servicios especializados, entre otros.

Modelos Neuronales de Lenguaje

El funcionamiento de los Grandes Modelos de Lenguaje se basa las redes neuronales artificiales, sistemas computacionales inspirados en el comportamiento del cerebro humano. Una red neuronal se compone por diferentes capas de neuronas conectadas entre sí; cada capa adquiere información de la capa anterior, la procesa y transmite información a la siguiente capa (Martin de Pablo, 2024).

Este método estableció los cimientos para la creación de modelos más avanzados en la interpretación de lenguajes.

Grandes Modelos de Lenguaje (LLM)

Los Grandes Modelos de Lenguaje (LLM) son tecnologías de inteligencia artificial que han sido capacitadas con extensos volúmenes de contenido textual, capaces de comprender y generar lenguaje humano de manera clara y relevante. Se basan en arquitecturas de redes neuronales profundas, particularmente en los Transformers, que emplean mecanismos de atención para identificar relaciones semánticas entre palabras y frases (Vaswani et al., 2017).

Según Shanahan (2024), los LLM son modelos generativos analizan la distribución estadística de fragmentos en amplios conjuntos de textos, incluyen palabras completas, partes de palabras o caracteres individuales como signos de puntuación y símbolos.

Para Martin de Pablo (2024), constituyen uno de los avances más significativos en IA, dado que permiten interactuar de manera fluida en lenguaje natural y abrir aplicaciones en chatbots, asistentes virtuales, traducción automática y sistemas de recuperación de información.

Modelos Representativos

GPT (OpenAI)

GPT es un modelo generativo que emplea transformadores y ha sido capacitado con extensos volúmenes de texto, lo que le da la capacidad de producir lenguaje natural de manera coherente (Yenduri et al., 2023).

Gemini (Google DeepMind)

Gemini es un modelo que ofrece múltiples modalidades y permite manejar contextos extensos, combinando texto, código, imágenes y sonido (Gemini Team, 2025).

Claude (Anthropic)

Se enfoca en fundamentos de ética y seguridad, incorporando una estructura que respeta los valores humanos en la comunicación dialogada (Anthropic, 2023).

LLaMA (Meta AI)

LLaMA es un modelo lingüístico accesible que ha sido creado con fines de investigación y está perfeccionado para operar en dispositivos con capacidades de hardware reducidas (Touvron et al., 2023).

Gemma (Google)

Gemma es un sistema lingüístico accesible de Google, diseñado para estudios y usos de inteligencia artificial generativa en diversos idiomas (Team Google Research, 2024).

DeepSeek

Es un modelo abierto que se enfocan en la eficiencia de la computación a través de arquitecturas de Mixture-of-Experts (MoE) y métodos de atención sofisticados como la

Atención Latente Multicabezal, lo que facilita la disminución de los costos de inferencia sin comprometer el rendimiento (DeepSeek-AI et al., 2025).

Análisis Comparativo

El avance de la Inteligencia Artificial ha fomentado la creación de Modelos de Lenguaje de Gran Escala (LLM), que son entrenados con enormes cantidades de información y emplean estructuras profundas como los Transformers. Estos modelos son reconocidos por su habilidad para producir y entender el lenguaje humano en diversas situaciones, siendo utilizados en aplicaciones como chatbots, asistentes digitales, categorización de textos y búsqueda de información (Vaswani et al., 2017; Brown et al., 2020).

Entre los LLM más destacados se pueden mencionar GPT de OpenAI, Claude de Anthropic, LLaMA de Meta, Gemma de Google y Gemini de Google DeepMind, cada uno con diversas características en cuanto a disponibilidad, licencias y usos. A continuación, se muestra cuadro comparativo que destaca las características relevantes de cada modelo:

Tabla 3

Comparativa de modelos LLM

Modelo	Características	Fortalezas	Limitaciones
GPT	Modelo generativo. Basado en transformadores.	Generación de lenguaje. Comunidad de soporte.	Dependencia de prompts. Costos elevados.
Gemini	Modelo multimodal (texto, código, imágenes, audio y video).	Capacidad de razonamiento. Optimizado para RAG.	Acceso aún restringido. Dependiente de infraestructura.

Modelo	Características	Fortalezas	Limitaciones
Claude	Enfocado en ética y seguridad. Prioriza comunicación.	Transparencia y mitigación de sesgos.	Menor multimodalidad. Centrado en texto.
LLaMA	Diseñado para investigación.	Gratuito y adaptable. Ideal para investigación.	Rendimiento inferior. Limitado en multimodalidad.
Gemma	Modelo multilingüe.	Soporte en múltiples idiomas.	Menor capacidad de razonamiento.
DeepSeek	Modelo de IA open source.	Abierto a la comunidad. Competitivo con modelos comerciales.	Menor madurez. Menos probado en producción.

Nota. Tomado de Anthropic (2023), DeepSeek AI (2024), Gemini Team (2025), Team Google Research (2024), Touvron et al. (2023), y Yenduri et al. (2023).

El modelo GPT (OpenAI) sigue siendo el líder en consistencia lingüística y apoyo, sin embargo, el costo y la falta de capacidades multimodales restringen su uso en organizaciones públicas. Por otro lado, Claude (Anthropic) se distingue por su enfoque ético y seguro, pero no se adapta a situaciones complejas. En cambio, LLaMA (Meta) y Gemma (Google) son valiosos para la investigación, pero no son adecuados para un sistema documental en funcionamiento. DeepSeek (China) tiene eficiencia y bajo costo, con una limitada madurez y soporte a nivel global. Es asi, Gemini (Google DeepMind) se halla como la alternativa destacada al combinar multimodalidad, razonamiento avanzado y optimización para RAG, asegurando contextualización y confiabilidad en tiempo real.

Para la selección del modelo se aplicó preparo la comparación de los **5 modelos LLM** usando una **escala de Likert de 1 a 5**, donde:

1 = Muy bajo

2 = Bajo

3 = Medio

4 = Alto

5 = Muy alto

Por ello, se valoró en **5 dimensiones:** Multimodalidad, Capacidad de razonamiento, Accesibilidad, Seguridad/ética y Rendimiento en producción.

Tabla 4

Comparativa de Scala de Likert

Modelo	Multimodalidad	Razonamiento	Accesibilidad	Seguridad / Ética	Rendimiento en producción	Total
GPT	2	4	3	3	4	16
Gemini	5	5	3	4	5	22
Claude	2	4	3	5	3	17
LLaMA	2	3	5	3	3	16
Gemma	3	3	4	3	3	16
DeepSeek	3	3	4	3	4	17

De acuerdo a la comparativa según la valoración Likert, Gemini es el modelo destacado con 22 puntos, gracias a su capacidad para manejar diferentes modos y su razonamiento superior. DeepSeek, que cuenta con 17 puntos, destaca por su eficiencia y rendimiento, pero todavía no alcanza a Gemini en lo que respecta a habilidades de producción masiva.

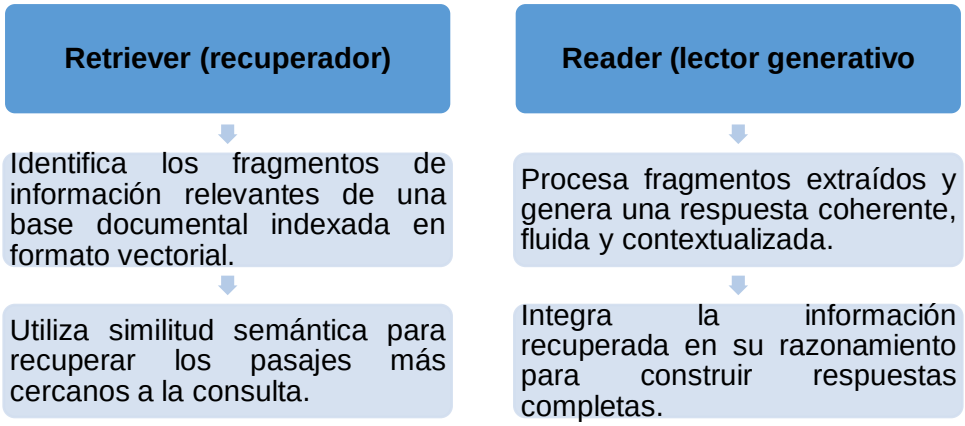
Retrevial – Augmented Generation (RAG)

Arquitectura

La arquitectura RAG (Retrieval-Augmented Generation) permite la recuperación de información semántica con la creación de texto en un contexto específico a través de Modelos de Lenguaje de Gran Escala (LLMs). Tiene como objetivo superar las limitaciones de los modelos generativos convencionales, que se basan únicamente en los datos que les fueron ofrecidos durante su formación, incorporando sistemas que facilitan la consulta en directo de bases documentales externas (Lewis et al., 2021). Este tipo de arquitectura posee dos componentes:

Figura 1

Componentes principales de la arquitectura Retriever–Reader



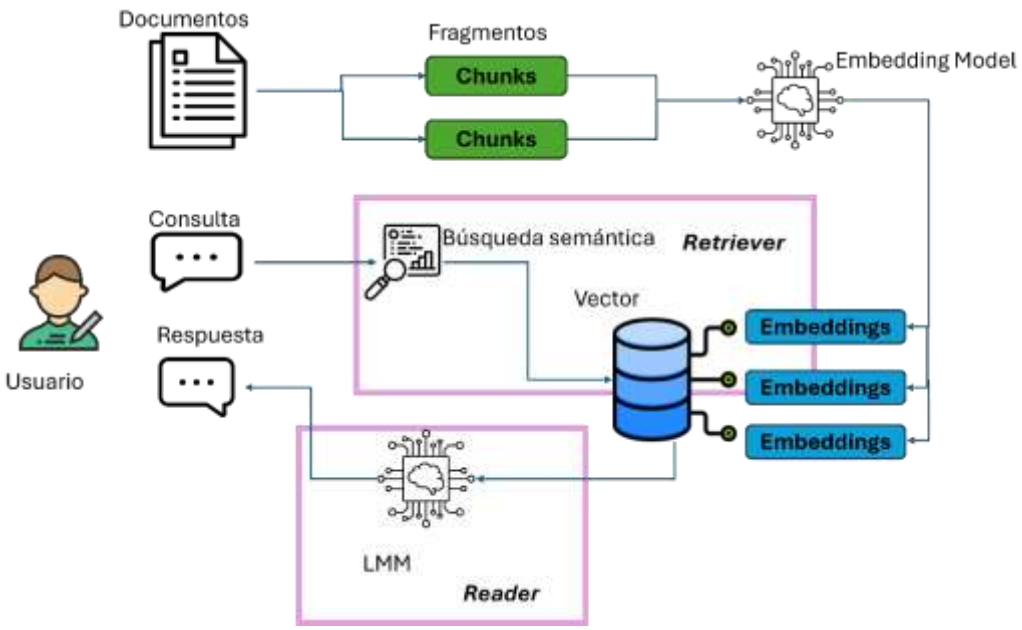
Nota. Tomado de Johnson, Douze y Jégou (2021) y Karpukhin et al. (2020).

Arquitectura de referencia

Como se puede observar en la siguiente figura, el Retriever examina los elementos más importantes en la base de datos vectorial (ChromaDB) basándose en la pregunta del usuario. Así también, el Reader (LLM) une esos elementos y produce una respuesta lógica y en contexto. De esta manera, la estructura RAG fusiona la búsqueda semántica y la creación para brindar respuestas exactas y fundamentadas.

Figura 2

Arquitectura de referencia RAG



La Figura 2 muestra una arquitectura de referencia del modelo RAG (Generación Aumentada por Recuperación) aplicado en el sistema inteligente de recuperación de documentos de EPAGAL. Este esquema fusiona la búsqueda semántica con la producción de lenguaje natural, uniendo los elementos Retriever y Reader para proporcionar respuestas acertadas y fundamentadas.

En la fase inicial, los documentos de la institución se fragmentan en partes y se convierten en representaciones en forma de vectores a través de un modelo de embeddings, que se guardan en ChromaDB como una base de datos semántica. Cuando se recibe una consulta en lenguaje natural, el módulo Retriever selecciona los fragmentos más pertinentes basándose en su similitud semántica, mientras que el Reader (LLM Gemini) analiza esa información y produce una respuesta coherente y contextual.

De esta manera, la estructura RAG combina la búsqueda semántica con la generación de lenguaje, facilitando respuestas fundamentadas y aumentando la eficacia en la recuperación de datos y el servicio al cliente.

Concepto y funcionamiento de embeddings

Los embeddings son representaciones numéricas de datos en un espacio vectorial continuo, y se utilizan en campo del procesamiento del lenguaje natural (PNL) para convertir palabras u otros elementos categóricos en vectores que capturan su significado semántico (Shi et al., 2023). Estas representaciones han revolucionado múltiples campos de la inteligencia artificial, ya que permiten que los algoritmos comprendan y gestionen datos textuales de manera más eficiente (Gonzáles Rivero, 2024). En el ámbito documental, los embeddings se utilizan para crear bases de datos en forma de vectores, donde el contenido extraído de los documentos se convierte en vectores y se guarda junto a información adicional como la ruta o el nombre del archivo.

Integración de EBR en RAG

La Recuperación Basada en Embeddings (EBR) se trata de obtener datos relevantes evaluando preguntas en relación con representaciones vectoriales de textos, fundamentándose en similitudes semánticas en lugar de en coincidencias exactas de términos (Lion, 2024). Este método resulta particularmente valioso en sistemas de búsqueda avanzada, en la automatización de respuestas a consultas y en la creación de contenido específico, donde es fundamental la exactitud en el contexto.

En este contexto, emerge la arquitectura RAG (Generación Aumentada por Recuperación), que integra la Búsqueda Basada en Evidencia con los Modelos de Lenguaje de Amplia Escala. El RAG opera en tres etapas:

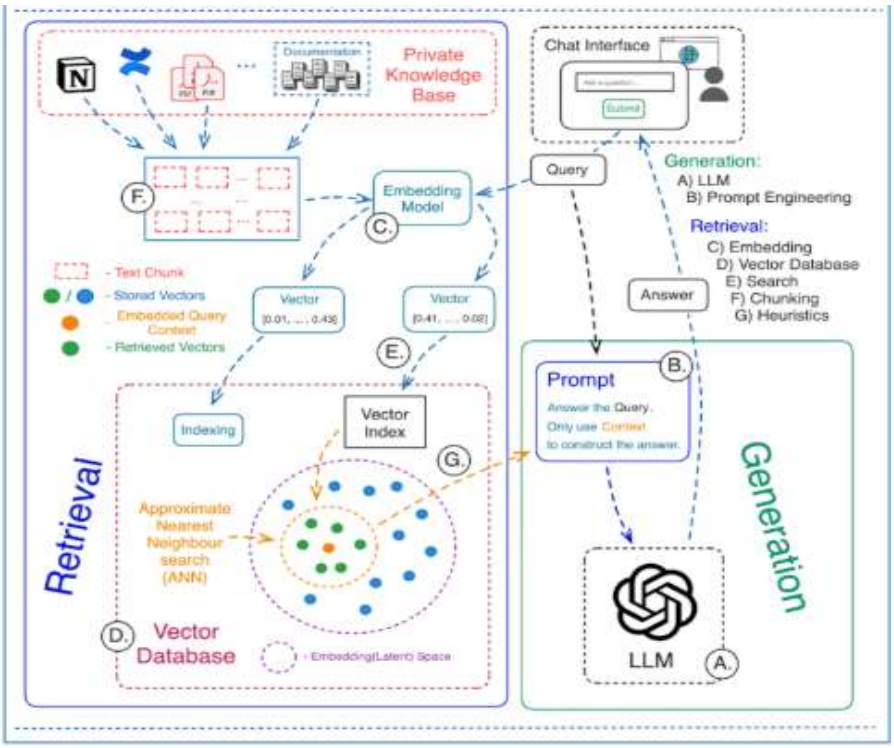
1. Convierte la consulta del usuario en un embedding.
2. Recupera fragmentos relevantes de la base vectorial mediante similitud semántica.
3. Envía los resultados al LLM, que los integra en su contexto y genera una respuesta enriquecida.

Por medio de este enfoque, el RAG proporciona respuestas que son precisas, coherentes y adaptadas al contexto, incluso si el modelo principal carece de información previa sobre el tema en cuestión. También utiliza sistemas de almacenamiento temporal

para preservar el contexto de la conversación y mejorar la efectividad del sistema (Huyen, 2022).

Figura 3

Integración EBR – RAG



Nota. Production grade Retrieval Augmented Generation (RAG) based LLM system. Tomado de Aurimas Griciūnas.

Bases de datos vectoriales

Las bases de datos vectoriales se emplean para almacenar, indexar y buscar vectores de alta dimensionalidad, cabe mencionar que es la representación numérica de datos textuales, imágenes y otros. Estas bases de datos se utilizan en aplicaciones en el cual los datos se realice el proceso de transformar en embeddings vectoriales, como puede ser la búsqueda semántica (Han et al., 2023).

Tabla 5

Cuadro comparativo de bases de datos vectoriales

Base de datos	Tipo / Licencia	Capacidades principales	Aplicaciones destacadas
OpenSearch	Open Source	Búsquedas de alto rendimiento. Indexación avanzada. Análisis en tiempo real. Detección de anomalías. Integración con ML.	Motores de búsqueda a gran escala, monitoreo de datos y sistemas analíticos.
PostgreSQL	Open Source	Almacena datos relacionales y JSON. Permite búsquedas vectoriales estructuradas + embeddings.	Aplicaciones RAG, IA. Gestión de datos estructurados y vectoriales.
ChromaDB	Open Source	Base de datos vectorial. Diseñada para IA generativa. Permite almacenar, indexar y consultar embeddings. Optimizada para integración con LLM y frameworks como LangChain.	Aplicaciones de RAG, chatbots con memoria contextual, búsqueda semántica en documentos.

Nota. Tomado de Esteban C (2023), Villarreal et al. (2017), Codeawake (2024) y Han et al. (2023).

En el análisis comparativo se puede concluir que, aunque OpenSearch y PostgreSQL tienen funciones importantes para la búsqueda y administración de datos, la opción que mejor se adapta al proyecto de EPAGAL es ChromaDB. Esto se debe a su diseño específico para aplicaciones de inteligencia artificial generativa, que permite una

conexión directa con herramientas modernas como LangChain, y que facilita la creación de sistemas de recuperación y generación (RAG) al mejorar el almacenamiento y la extracción de embeddings. Su característica de ser ligera, de código abierto y enfocada en la búsqueda semántica la hacen la alternativa más efectiva para la gestión de documentos y atención al cliente en contextos públicos, tal como se presenta en este estudio.

Por ello, se realizó un análisis comparativo con escala de Likert de 1 - 5, donde el modelo ganador es ChromaDB, ya que está optimizado para RAG y LLM:

Tabla 6

Comparativa con escala Likert

Criterio	OpenSearch	PostgreSQL	ChromaDB
Facilidad de integración con LLM	3	4	5
Especialización en búsquedas vectoriales	3	4	5
Escalabilidad y rendimiento	5	4	4
Versatilidad (estructurado + vectorial)	3	5	4
Enfoque en RAG y aplicaciones IA	3	4	5

Como se puede observar la Tabla 6, ChromaDB obtiene la calificación más alta (23/25), consolidándose como la opción preferida para arquitecturas RAG y aplicaciones que utilizan modelos de lenguaje por su especialización en embeddings, fácil integración y enfoque en inteligencia artificial generativa. PostgreSQL, incluyendo extensiones como PgVector, consigue un buen equilibrio entre flexibilidad y soporte híbrido (21/25), siendo perfecto para situaciones donde se necesitan datos tanto estructurados como vectoriales en un solo sistema. En cambio, OpenSearch (17/25) se destaca en escalabilidad y supervisión de datos, pero no logra el nivel de optimización en inteligencia artificial y RAG que ofrece ChromaDB.

Agentes Conversacionales

Son plataformas digitales que se comunican con las personas a través de un lenguaje comprensible, simulando una charla humana con diversos niveles de dificultad (Allouch et al., 2021).

Este tipo de agentes integran comprensión del lenguaje, entendimiento del idioma, manejo de conversaciones y creación de respuestas, facilitando apoyo en áreas como servicios, educación o sanidad (Sadek et al., 2023).

IA Generativa

Es una disciplina dentro de la inteligencia artificial que genera contenido nuevo, ya sea en forma de texto, imágenes, sonido o programación, basándose en datos previos y empleando modelos sofisticados como LLMs y GANs (Park et al., 2023).

Su importancia se encuentra en la habilidad de producir datos que son lógicos y relevantes, con usos en enseñanza, creación, estudio e industria pública (Sidaoui et al., 2024).

Bajo este contexto, la IA Generativa se encarga de crear contenido de diferente tipo, sea de texto como ChatGPT y Gemini, imágenes, videos, música entre otros.

Arquitecturas de diálogo

Son los marcos que organizan la interacción entre el cliente y el sistema, incluyendo aspectos de comprensión, gestión y generación de respuestas (Huq et al., 2024).

Se desarrollan desde sistemas fundamentados en reglas hasta estructuras híbridas que combinan aprendizaje profundo con memoria contextual (Liu et al., 2024).

Las arquitecturas de diálogo han cambiado notablemente desde los primeros programas que utilizaban reglas, como ELIZA en los años sesenta, que solo replicaban charlas simples, hasta los modelos actuales sustentados en LLM que pueden mantener interacciones más elaboradas y con más contexto (WeizenbaumJoseph, 1983). Existen

arquitecturas relevantes como seq2seq, que introdujeron la traducción automática de entradas en salidas, los transformers, que mejoraron la capacidad de atención y escalabilidad, y los esquemas RAG (Retrieval-Augmented Generation), ayudan a recuperar información con generación de texto, complementados con mecanismos de memoria a corto y largo plazo para dar continuidad a los diálogos (Lewis et al., 2020; Vaswani et al., 2017).

Asistentes Inteligentes

Son sistemas impulsados por inteligencia artificial que ayudan al usuario a completar tareas usando lenguaje natural y adaptación al contexto (Allouch et al., 2021).Un agente inteligente se define como una entidad (ya sea física o virtual) que tiene la capacidad de observar su entorno a través de sensores, interpretar esas observaciones para reflexionar (concluir, razonar, tomar decisiones) y actuar sobre el entorno utilizando actuadores con el fin de lograr metas específicas de manera independiente. Además, suele tener la habilidad de aprender de sus experiencias para optimizar su rendimiento (Cheng et al., 2024).

Los asistentes inteligentes se categorizan de diferentes formas según su función y área de uso. Estos son asistentes personales enfocados en ayudar a las personas con actividades diarias como recordatorios o el manejo de calendarios, los asistentes educativos cuyo objetivo es facilitar el aprendizaje mediante tutorías, evaluaciones, comentarios personalizados y los asistentes empresariales, que colaboran en la administración de relaciones con clientes, ventas en línea y procedimientos internos.

Human-in-the-loop (HITL)

Es un método que combina la participación humana en el proceso de formación y evaluación de modelos, aumentando su exactitud y disminuyendo prejuicios (Mosqueira-Rey et al., 2023).Proporciona claridad y supervisión humana en el proceso de decisiones basadas en algoritmos, lo cual es fundamental en sistemas importantes (Park et al., 2023).

El método Human-in-the-loop (HITL) se utiliza en varios entornos prácticos. Ejemplos más frecuentes puede ser el etiquetado de información, donde las personas

ajustan o verifican las categorías que han sido asignadas por algoritmos esto contribuye a mejorar la calidad del entrenamiento de los modelos (Shneiderman, 2020) (Mosqueira-ReyEduardo et al., 2022). Sin embargo, el enfoque presenta limitaciones y críticas como el costo en tiempo y recursos humanos puede ser elevado, especialmente en proyectos que requieren gran volumen de datos.

Aplicaciones y uso de IA en la Administración Pública

La inteligencia artificial se utiliza en el ámbito gubernamental para optimizar la gestión administrativa, examinar grandes cantidades de información y brindar servicios más adaptados a las necesidades de los ciudadanos (Madan & Ashok, 2023).

Su adopción permite optimizar políticas públicas, incrementar la transparencia y fortalecer la participación ciudadana (Babšek et al., 2025).

En el ámbito de la gestión pública y servicios, los Modelos de Lenguaje de Gran Escala (LLM) y las técnicas de Recuperación Basada en Embeddings (EBR) se ha iniciado a aplicarse para mejorar la gestión documental y atención al cliente. Por ello, diversas entidades públicas, como municipios u organismos gubernamentales en Europa y América Latina han incorporado chatbots inteligentes para responder consultas normativas, procesar solicitudes en línea y facilitar el acceso a reglamentos y procedimientos (Zhou et al., 2024).

Cabe mencionar, en el sector ambiental, se han creado sistemas de información que incorporan inteligencia artificial para clasificar leyes, reportes técnicos y normativas, facilitando a los empleados el acceso a los datos de manera rápida y precisa. (Lima, 2024). Empresas de servicios como Aguas de Barcelona o la Agencia Ambiental de California han comenzado a utilizar soluciones de recuperación semántica basadas en IA para mejorar la trazabilidad de documentos internos y dar soporte a la ciudadanía.

Estos avances indican de qué manera la inteligencia artificial está cambiando la administración de documentos en el sector público y en los servicios. Esto representa una

base importante para implementar estas tecnologías en EPAGAL, donde se hace cada vez más clara la necesidad de optimizar la recuperación de normas y procedimientos internos.

Tecnologías para el sistema propuesto

Lenguaje de Programación de Backend y Frontend

Es un lenguaje de programación orientado al desarrollo de aplicaciones relacionadas con la Inteligencia Artificial y el procesamiento de información. Esto se debe por su robusto ecosistema, rápida adaptabilidad y evolución con librerías modernas (Alfaro & Roberto, 2022).

Python se distingue por ofrecer gran variedad de bibliotecas especializadas para Inteligencia Artificial (IA) y Procesamiento de Lenguaje Natural (NLP) (Reimers & Gurevych, 2019), estos ayudan a la generación de embeddings semánticos; como también frameworks para orquestación como RAG, y servidores web de gran rendimiento como FastAPI. Lo antes mencionado convierte a Python el núcleo del sistema inteligente propuesto, este es el lenguaje responsable de coordinar la lógica del RAG (Embedding Based Retrevial + LLM), además de coordinar las operaciones Backend que dan soporte al chatbot.

Framework para Implementación de RAG

LangChain

LangChain se ha establecido como un marco de trabajo de código abierto que simplifica el desarrollo de aplicaciones fundamentadas en Modelos de Lenguaje de Gran Tamaño (LLM), al combinar de manera efectiva embeddings, bases de datos vectoriales y secuencias modulares, lo que permite enlazar los modelos de lenguaje con fuentes de datos externas (DigitalOcean, 2024).

Bajo este contexto, se presenta una herramienta nueva que proporciona un método flexible y adaptable para la creación de aplicaciones, facilitando las etapas de diseño,

elaboración e integración contextual —sumando la Recuperación Aumentada de Generación (RAG)— y ofreciendo funciones relacionadas con el estado, el contexto y la seguridad en las interacciones (Mavroudis, 2024). De este modo, LangChain no solo simplifica la complejidad técnica, sino que también incrementa las aplicaciones de los LLM en contextos de trabajo e investigación, estableciéndose como un elemento fundamental en la coordinación de sistemas inteligentes.

El framework está diseñado para la gestión de pipelines RAG (Generación Aumentada por Recopilación), lo que hace más sencillo crear aplicaciones inteligentes que integran LLM con bases de datos en forma de vectores. Su empleo ofrece:

Este método de trabajo amplía de manera significativa las posibilidades de uso práctico de los LLM, abarcando desde asistentes digitales hasta plataformas de sugerencias. Además, lo más relevante es que pueden ofrecer datos a los LLMs que, debido a su formación, previamente no tenían. (*LangChain Docs Home*, s. f.).

Desarrollar indicaciones mejoradas que integren la información recuperada con la pregunta inicial del usuario, mejorando así la producción del modelo de lenguaje.

Añadir características extra como la supervisión de metadatos, para garantizar que se pueda rastrear el origen de los fragmentos utilizados, así como establecer procesos de respaldo, en situaciones donde no haya contexto suficiente para crear una respuesta fiable. (*LangChain State of AI 2023*).

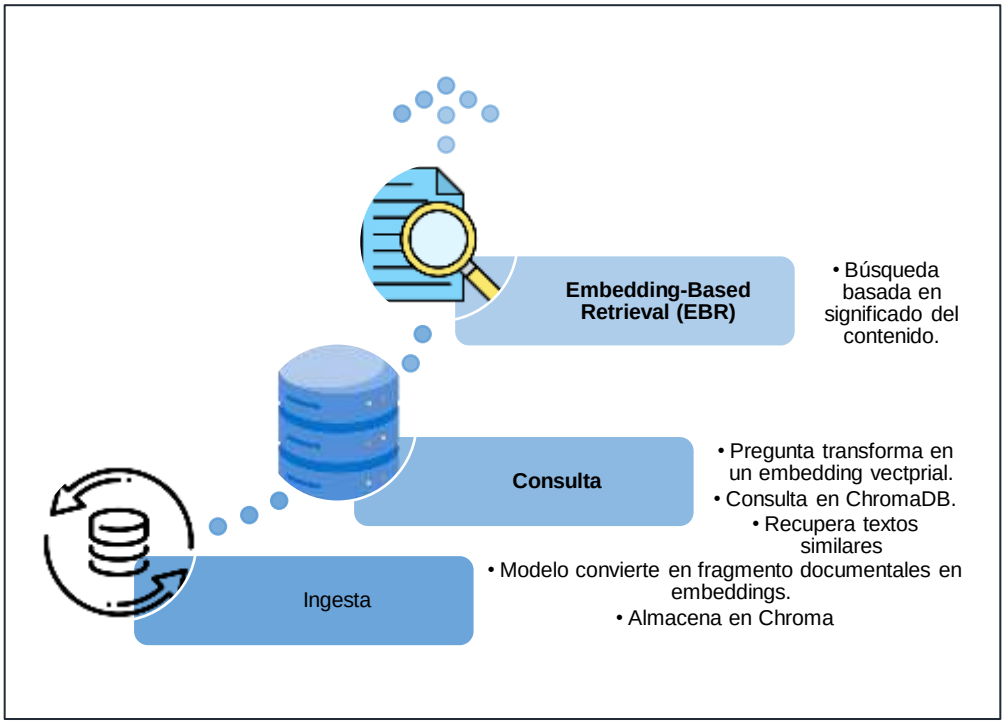
Embedding con Sentence-Transformers

Se considera una técnica de procesamiento de lenguaje natural que convierte el lenguaje humano en vectores matemáticos. Estos vectores representan el significado fundamental de las palabras, facilitando así que las computadoras manejen el lenguaje de forma más eficiente (Reimers & Gurevych, 2019).

En este proyecto se, emplean el modelo all-MiniLM-L6-v2 que pertenece a los sentence-transformers, cuya finalidad es producir embeddings de oraciones o fragmentos de texto. (Reimers & Gurevych, 2019)

Figura 4

Funcionamiento de Embedding-Based Retrieval (EBR)



Nota. Tomado de Reimers & Gurevych (2019) y Sentence-Transformers (2024).

La Figura 3, muestra el proceso de ingesta, consulta y recuperación semántica por medio de Embedding-Based Retrieval (EBR), utilizando modelos de sentence-transformers como all-MiniLM-L6-v2 para representar documentos y consultas en un espacio vectorial semántico.

En el sistema de recuperación documental inteligente de EPAGAL, el EBR tiene un papel fundamental en la etapa intermedia del proceso RAG (Generación Aumentada por Recuperación):

En el proceso de ingestión de documentos, transforma los archivos administrativos en vectores y los almacena en ChromaDB, conservando su estructura semántica.

Al interactúa con el usuario, convierte las preguntas en embeddings y lleva a cabo la búsqueda de similitudes.

Los resultados obtenidos se envían al modelo de lenguaje LLM Gemini, que produce una respuesta contextual basada en los documentos de la institución.

Base vectorial ChromaDB

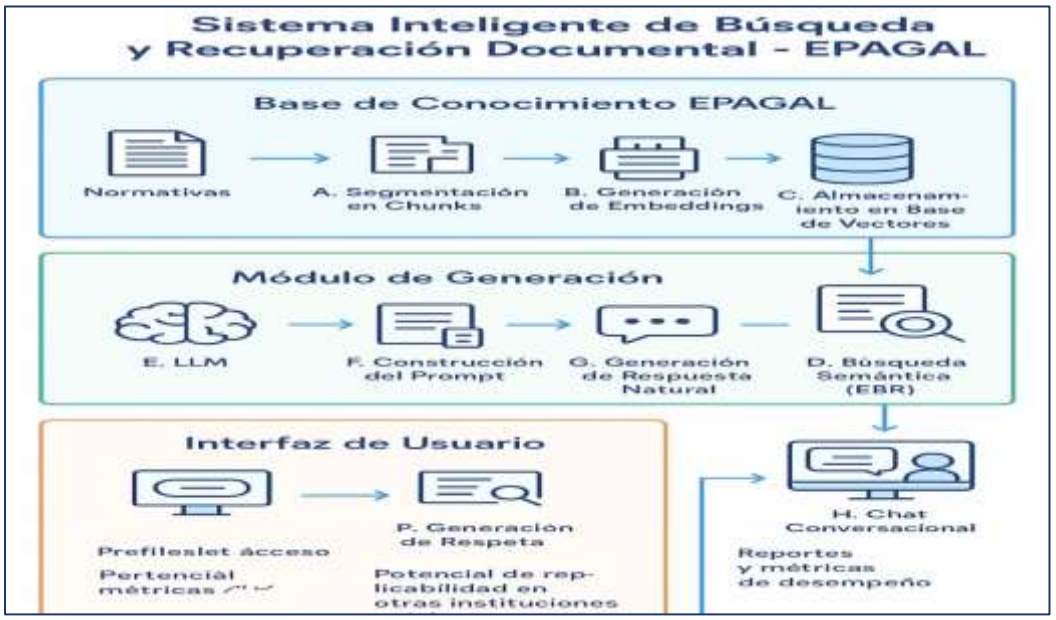
Es una base de datos vectorial de código abierto, que ayuda a la creación de aplicaciones LLM, el objetivo es almacenar vectores de embeddings y los documentos asociados.

En el sistema propuesto, ChromaDB cumple la función de:

- Guardar los embeddings creados a partir de documentos durante la etapa de ingesta.
- Realizar la recuperación semántica, devolviendo los fragmentos más relevantes que se utilizarán como contexto para el LLM (Chroma, s. f.).

Figura 5

Sistema Inteligente de Búsqueda y Recuperación Documental EPAGAL



Capitulo III

Propuesta Metodológica y Diseño de Solución de Búsqueda de Información Documental

Integración de RAG y Gestión Documental

La combinación de la arquitectura RAG (Generación Aumentada por Recuperación) con los procedimientos de gestión documental de EPAGAL constituye un pilar en la metodología propuesta. Mientras la gestión documental garantiza la sistematización y conservación de la información normativa, el modelo RAG facilita la conversión de estos archivos en sistemas de consulta inteligentes. De tal forma, que el ciudadano como el funcionario no solo tienen acceso a documentos completos, sino que reciben respuestas contextualizadas, basadas en los fragmentos significativos del conjunto documental institucional. A continuación, se describe de forma textual el proceso:

En EPAGAL, la gestión de documentos se realiza mediante el tratamiento de normativas, regulaciones, contratos, informes técnicos y archivos administrativos. Hoy en día, los registros se guardan en formato digital como PDF y Word en repositorios internos. No obstante, uno de los principales inconvenientes es la poca eficiencia en la localización de información, ya que depende de palabras clave precisas o de la búsqueda manual del documento, lo que complica el acceso ágil a contenidos concretos y provoca retrasos tanto para los empleados como para los ciudadanos.

Esta propuesta incluye un sistema de generación aumentada por recuperación (RAG) que se apoya en tecnologías actuales como LangChain para la gestión, ChromaDB como base de datos vectorial, y Gemini como modelo de lenguaje de gran tamaño (LLM). El proceso abarca tres etapas: la recopilación de documentos normativos (que se transforman en fragmentos y representaciones vectoriales), las consultas de los usuarios en lenguaje natural (que se convierten en vectores), y la búsqueda semántica de fragmentos relevantes

para ofrecer una respuesta contextual. Este sistema está dirigido tanto a empleados públicos, quienes necesitan acceso rápido a los procedimientos internos, como a ciudadanos, que desean información normativa de forma clara y sencilla.

La combinación se lleva a cabo en el momento de la recopilación de documentos, en el cual los archivos de la institución, como las resoluciones en formato PDF, son transformados en trozos y luego se representan como vectores utilizando modelos de Sentence-Transformers. Estos vectores se guardan en ChromaDB, que funciona como el repositorio semántico adicional al archivo documental convencional. De este modo, RAG actúa como la capa de inteligencia para la búsqueda, mientras que la gestión documental continúa desempeñando su función en la conservación, clasificación y supervisión de los documentos oficiales. Para garantizar la trazabilidad, cada parte indexada en ChromaDB contiene metadatos que permiten seguir el documento original y su versión asociada.

Se utiliza un agente conversacional como apoyo en los procesos de análisis, redacción y revisión de información. El uso de este tipo de herramienta se fundamenta en su capacidad para simular una interacción humana, ofreciendo respuestas contextualizadas y adaptadas a las necesidades del usuario.

El empleo de un agente conversacional resulta útil porque facilita la búsqueda, organización y comprensión de la información, permitiendo optimizar el tiempo y mejorar la claridad en la elaboración de los contenidos. Más allá de ser una herramienta tecnológica, representa un medio de apoyo académico que contribuye al desarrollo del pensamiento crítico y al fortalecimiento de las habilidades comunicativas.

En este trabajo, el agente conversacional fue utilizado como asistente para estructurar ideas, revisar coherencia textual y verificar el cumplimiento de normas académicas, sin sustituir el criterio personal ni el análisis propio del usuario. Su función se limitó a complementar la labor humana, ofreciendo sugerencias y orientaciones que posteriormente fueron revisadas, interpretadas y adaptadas de forma consciente.

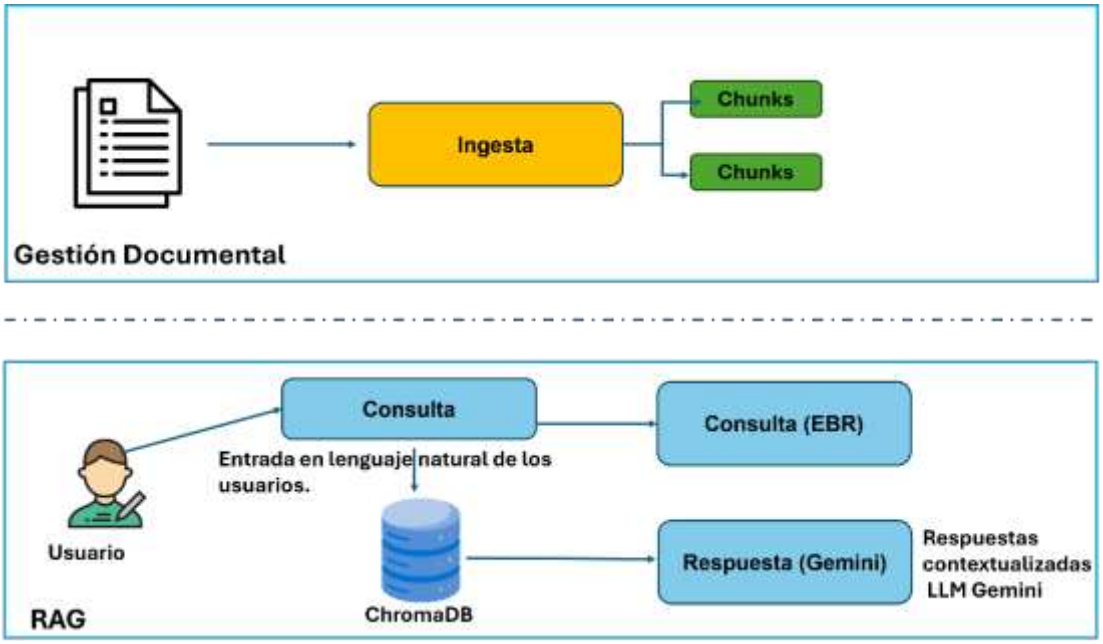
En términos generales, el uso de este tipo de tecnología demuestra cómo la inteligencia artificial puede integrarse de manera responsable en el ámbito académico y

administrativo, potenciando la productividad sin perder la esencia del pensamiento reflexivo y creativo del investigador.

A continuación, se ilustra cómo se combina la gestión de documentos con RAG, en el que los documentos se dividen en embeddings guardados en ChromaDB. Luego, el usuario hace una pregunta en un lenguaje sencillo y el LLM Gemini produce respuestas relevantes basadas en la recuperación semántica.

Figura 6

Integración entre gestión documental y RAG



La Figura 4, a integración entre la gestión documental y la arquitectura RAG (Generación Aumentada por Recuperación) aplicada en el sistema inteligente de búsqueda documental de EPAGAL. Los documentos oficiales son ingeridos, procesados y fragmentados en *chunks*, los cuales se transforman en embeddings y se almacenan en ChromaDB como repositorio semántico.

El usuario realiza consultas en lenguaje natural que son procesadas por el módulo EBR, el cual busca en ChromaDB los fragmentos relevantes. Luego, el LLM Gemini genera una respuesta contextualizada que se muestra en la interfaz conversacional. Esta

integración permite ofrecer respuestas precisas y verificables, fortaleciendo la gestión documental y mejorando la atención al cliente en EPAGAL.

Caso práctico: Empresa Pública de Aseo y Gestión Ambiental del Cantón Latacunga – EPAGAL

El objetivo del sistema sugerido es proporcionar a los usuarios respuestas en un lenguaje natural, relevantes y basadas en los documentos oficiales de EPAGAL, sin que se requiera tener un entendimiento previo de la organización o el formato de esos documentos. De esta manera, se pretende facilitar el acceso de los ciudadanos a la información normativa y operativa de la institución.

Arquitectura Propuesta

En la arquitectura propuesta en la Figura 5, se muestra el RAG integrada con gestión documental, aquí los documentos se vectorizan y se almacenan en ChromaDB junto con metadatos para ello el usuario realiza una consulta en lenguaje natural que se vectoriza y pasa por la aplicación orquestadora en Python; esta consulta activa el proceso de Embedding-Based Retrieval en ChromaDB, recuperando fragmentos relevantes que se envían al LLM (Gemini), el cual genera una respuesta contextualizada para el usuario.

Figura 7

Arquitectura Sistema Inteligente



La arquitectura planteada es un sistema inteligente basado en RAG (Retrieval-Augmented Generation), compuesto por dos elementos:

- EBR (Embedding-Based Retrieval) permite localizar documentos relevantes mediante de similitud semántica.

- LLM (Modelo de Lenguaje Extenso) permite a generar una respuesta natural y contextualizada basada en información recuperada.

El sistema localiza las secciones relevantes en la base de datos vectorial y, a partir de eso, el LLM genera una respuesta clara y basada en fundamentos.

Componentes de la arquitectura

1. **Vectorización de contexto**

- Para que el sistema pueda responder a preguntas, se organiza un conjunto de documentos como archivos de PDF, Word, Txt, entre otros.

- Cada fragmento se convierte en un vector semántico gracias al modelo Sentence-Transformers, un proceso que se denomina vectorización.

2. Inserción de la base vectorial (Chroma DB)

Los embeddings generados con los metadatos se insertan en Chroma DB, esta base de datos está orientada a búsquedas por similitud.

La elección de ChromaDB se fundamenta en rendimiento, escalabilidad, integración con LangChain, y la orientación a búsquedas semánticas como se puede observar en la Tabla 7:

Tabla 7

Elección de ChromaDB como base vectorial

Criterio	ChromaDB
Rendimiento y escalabilidad.	Diseñada para manejar millones de vectores de alta dimensionalidad. Garantiza tiempos de respuesta reducidos en consultas masivas.
Integración con frameworks modernos.	Ofrece conectores nativos con librerías de orquestación como LangChain. Utilizadas en arquitecturas RAG.
Compatibilidad con múltiples formatos de datos.	Almacena embeddings provenientes de texto, documentos normativos, imágenes o audio.
Orientación a búsquedas semánticas.	Recupera fragmentos relevantes. Superando las limitaciones de bases relacionales o NoSQL tradicionales.

Nota. Tomado de Han et al. (2023), Lewis et al. (2021), García & Rivera (2024), y Negueruela Gómez (2024).

En este sentido, ChromaDB se posiciona como la opción más adecuada para EPAGAL, ya que posibilita implementar un sistema de recuperación documental capaz de responder en lenguaje natural a las preguntas ciudadanas, con rapidez, precisión y pertinencia.

3. **Consulta del usuario**

Los usuarios interactúan con el LLM enviado consultas, a través de una interfaz web o móvil. La consulta se hace en una pregunta en lenguaje natural.

El usuario sea ciudadano o funcionario público, se conecta al sistema mediante una interfaz web creada con React, donde puede hacer preguntas en lenguaje natural sobre documentos oficiales, regulaciones o trámites administrativos.

La parte frontal del sistema se enlaza con el backend desarrollado en FastAPI, que recibe la pregunta y la envía al módulo orquestador programado en Python (LangChain).

Este módulo se encarga de convertir la solicitud en texto a una forma computacional adecuada para su manejo posterior.

4. **Búsqueda del contexto en la base vectorial**

Una vez recibida la pregunta, el orquestador (LangChain + Python) lleva a cabo el procedimiento de Recuperación Basada en Embeddings (RBE), que incluye las siguientes actividades:

La pregunta en formato de texto se transforma en un vector numérico (embedding) utilizando un modelo de sentence-transformers (como all-MiniLM-L6-v2).

El vector producido se emplea para llevar a cabo una búsqueda de similitud semántica en la base de datos vectorial ChromaDB, aplicando el método de Vecinos Más Cercanos Aproximados (ANN).

ChromaDB evalúa el espacio vectorial y recupera los fragmentos de texto que están más próximos al significado de la pregunta, enfatizando la intención y el contexto, en lugar de enfocarse únicamente en coincidencias de palabras.

Ejemplo:

- **Consulta:** "¿Cuáles son las etapas para manejar desechos sólidos?"
- **Recuperación:** fragmentos de normativa sobre residuos y desechos orgánicos.
- **Respuesta del LLM:** "Las etapas para el tratamiento de desechos sólidos incluyen la clasificación, recolección, transporte, tratamiento y disposición final."

5. **Devolución del contexto**

Una vez que ChromaDB detecta los fragmentos significativos, el sistema envía esos resultados (fragmentos) al módulo orquestador. El backend en Python combina la consulta original del usuario con los fragmentos recuperados, estructurando un prompt extendido.

El prompt contiene el contexto documental preciso para que Gemini como modelo de lenguaje comprenda el tema y genere una respuesta fundamentada.

El formato del prompt puede incluir delimitadores o instrucciones explícitas para asegurar que el LLM responda únicamente basándose en la información documental obtenida.

6. **Envió al LLM**

En esta fase el prompt extendido se envía al modelo de lenguaje LLM Gemini, este actúa componente generativo del sistema.

El modelo:

- Recibe la consulta y el contexto integrados en el *prompt*.
- Procesa la información mediante sus redes neuronales entrenadas en lenguaje natural.
- Interpreta la intención del usuario y genera una respuesta en lenguaje natural, coherente y respaldada por los documentos institucionales.

7. Respuesta generada

El modelo del LLM devuelve una respuesta en lenguaje natural, centrada en la información encontrada en los documentos institucionales, la cual se visualiza en la interfaz conversacional del sistema.

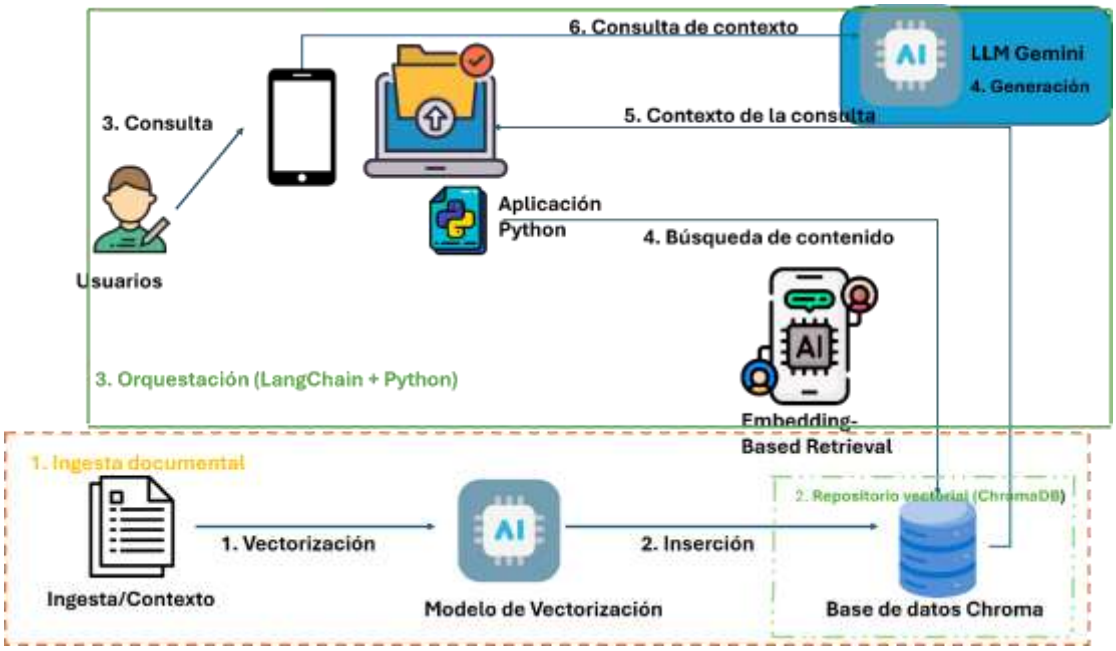
El modelo LLM Gemini devuelve una respuesta que es generada al backend, donde la aplicación Python/FastAPI la formatea y envía al frontend React y muestra al usuario que visualiza en pantalla.

Diseño del modelo de búsqueda documental

El modelo de búsqueda de documentos se basa en la estructura RAG (Generación Aumentada por Recuperación), que une la recuperación semántica de partes de documentos con la creación de respuestas contextualizadas por un modelo de lenguaje de gran tamaño (LLM). Esta configuración tiene como objetivo superar las restricciones de los sistemas clásicos que dependen de palabras clave, ofreciendo respuestas que son precisas, fundamentadas y conectadas a la documentación oficial de EPAGAL. A continuación, se identifica cada una de las capas del modelo:

Figura 8

Mapeo de capas del modelo



En la Figura 8, se observa la estructura de capas del modelo para la búsqueda de documentos, fundamentado en la arquitectura RAG (Generación Aumentada por Recuperación). Este método integra la búsqueda semántica de partes documentales con la creación contextual de respuestas utilizando un modelo de lenguaje de gran tamaño llamado LLM Gemini.

Con este diseño es busca resolver las limitaciones que tienen los sistemas convencionales basados en palabras clave, permitiendo proporcionar respuestas precisas, bien fundamentadas y relacionadas con la documentación oficial de EPAGAL.

El modelo se compone de **cuatro niveles funcionales** que funcionan de forma secuencial y se apoyan mutuamente:

1. **Ingesta documental**
    - Los archivos normativos y administrativos (PDFs, resoluciones, informes) son fragmentados en *chunks* y convertidos en embeddings.

    - Este método clasifica la información en formas que ayudan en búsquedas futuras.

2. Repositorio vectorial (ChromaDB)

- Los embeddings generados se almacena en Chroma junto con metadatos (fuente, fecha, versión) para asegurar trazabilidad.

- Esta estructura vectorial facilita encontrar similitudes en el significado, en vez de buscar coincidencias que sean textualmente precisas.

3. Orquestación LangChain + Python

La capa de orquestación forma el corazón funcional del sistema inteligente de búsqueda de documentos, puesto que se encarga de coordinar la interacción entre la interfaz del usuario, el repositorio de vectores (ChromaDB) y el modelo de lenguaje LLM (Gemini).

Su puesta en marcha se lleva a cabo a través del marco LangChain, que funciona como un enlace flexible entre los diferentes elementos de IA, y el lenguaje Python, que se ocupa de realizar las operaciones lógicas, consultas y el manejo del contexto.

Tabla 8

Proceso de orquestación LangChain + Python

Etapa	Descripción	Componente principal
1. Recepción de la consulta	El usuario formula una pregunta en lenguaje natural y Python la envía a LangChain como *prompt*.	Interfaz web – Python
2. Vectorización de la	LangChain convierte la consulta en un vector	LangChain +

consulta	semántico mediante un modelo de *embeddings*.	Embeddings
3. Recuperación EBR	Se busca en **ChromaDB** los fragmentos más similares según la intención del usuario.	LangChain + ChromaDB
4. Construcción del contexto	Python integra la consulta y los fragmentos recuperados en un *prompt* extendido.	Python + LangChain
5. Interacción con el LLM	El *prompt* se envía al modelo **Gemini**, que genera una respuesta coherente y fundamentada.	LangChain + LLM Gemini
6. Respuesta al usuario	La respuesta final se muestra en la interfaz conversacional, con opción de citar las fuentes.	Python + Interfaz web/móvil

4. Generación (LLM Gemini)

Produce respuestas claras, fundamentadas y contextualizadas utilizando los fragmentos recuperados como base de conocimiento.

Dentro del sistema inteligente de EPAGAL, la capa LLM Gemini actúa como el mecanismo de síntesis y comunicación entre el conocimiento institucional y el usuario final.

Funcionamiento

- **Recepción del contexto ampliado.** - El modelo LLM Gemini recibe un mensaje mejorado que incluye la consulta inicial del usuario, los fragmentos obtenidos de ChromaDB a través del procedimiento EBR. Este mensaje forma el contexto sobre el cual el modelo desarrolla su respuesta.
- **Procesamiento semántico. -** Gemini analiza el texto por medio de sus redes neuronales de atención y comprensión del lenguaje, identificando las relaciones

entre la consulta y la información contextual. Determina la importancia de cada sección para eliminar datos que sean repetitivos o irrelevantes.

- **Generación de respuesta. -** El modelo Gemini crea una respuesta fluida y natural, y prioriza la claridad del lenguaje comprensible para usuarios administrativos y ciudadanos. Y contextualización que hace referencia implícita al marco institucional (normas, procesos o informes).
- **Entrega al sistema orquestador. -** La respuesta generada se envía de vuelta al módulo Python + LangChain, que la formatea y muestra en la interfaz conversacional.

En esta capa garantiza que el sistema no solo obtenga datos, sino que también analice y transmita información, proporcionando a EPAGAL un recurso seguro para el servicio al ciudadano y la administración de documentos de manera inteligente.

Capítulo IV
Implementación de la Propuesta de Solución

Fases de Desarrollo

Fase 1: Preprocesamiento y Vectorización Documental

Los documentos oficiales de EPAGAL, que incluyen normas, reglamentos, manuales y procedimientos, fueron reunidos en formato PDF, que se estableció como norma debido a su estabilidad y capacidad para mantener la estructura. Se utilizaron las bibliotecas PyMuPDF (fitz) y pdfminer. six para extraer el contenido de texto, lo que facilitó la obtención de un texto limpio y la eliminación de elementos innecesarios como tablas e imágenes.

Código:

```
base_conocimiento = []
pdf_files = [f for f in os.listdir(PDF_FOLDER) if f.endswith(".pdf")]

id_counter = 1

for pdf_file in pdf_files:
    full_path = os.path.join(PDF_FOLDER, pdf_file)
    print(f"Procesando: {pdf_file}")

    # Intenta extraer texto
    text = extract_text_from_pdf(full_path)

    if len(text.strip()) < 100:
        print("➤ Texto insuficiente, usando OCR...")
        text = ocr_pdf(full_path)

    # Divide en bloques de hasta 1000 caracteres
    bloques = [text[i:i+1000] for i in range(0, len(text), 1000)]

    for bloque in bloques:
```

```
        prompt = plantilla.format(
            contenido=bloque,
            source=pdf_file,
            fecha_actual=hoy
        )
        try:
            respuesta = llm.invoke(prompt)
            data = json.loads(respuesta)
            data["id"] = str(id_counter)
            base_conocimiento.append(data)
            id_counter += 1
        except openai.error.RateLimitError as e:
            print(f"Límite de cuota excedido: {e}")
            # Opcional: salir del programa para no seguir haciendo llamadas
            break
        except Exception as e:
            print(f" Error en bloque del archivo {pdf_file}: {e}")
            print("Respuesta del modelo:\n", respuesta)
```

Fase 2: Limpieza y Segmentación de Fragmentos.

El contenido que se obtiene se organiza en secciones de un máximo de 1000 caracteres, asegurando que cada parte disponga de datos suficientes sin exceder los límites permitidos por el modelo generativo (GPT-3. 5/4). Esta división permite una indexación y recuperación más efectiva.

Código:

```
bloques = [text[i:i+1000] for i in range(0, len(text), 1000)]
```

Fase 3: Vectorización Semántica (LangChain)

Creada la base de conocimiento, con los fragmentos textuales procesados y anotados por LLM, estos fragmentos se convierten en vectores semánticos. El proceso de

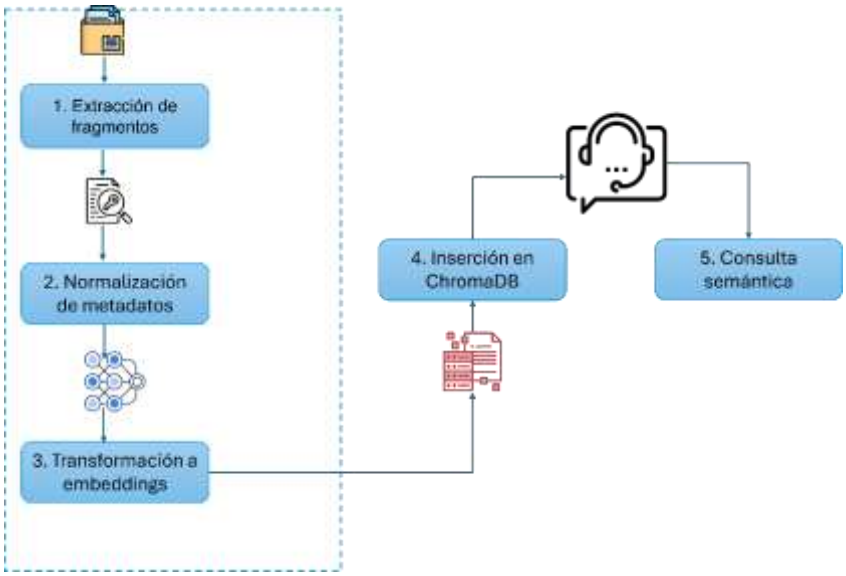
transformación representa en contenido textual en un espacio de muchas dimensiones, lo cual hace más sencilla la búsqueda basada en similitudes conceptuales.

En esta sección se emplea el modelo preentrenado all-MiniLM-L6-v2 de la biblioteca de Sentence-Transformers, este es utilizado por la precisión semántica, eficiencia y velocidad.m Para el desarrollo del proyecto, se utiliza el modelo preentrenado all-Mini-L6-v2 proporcionada por la biblioteca sentence-transformers, útil para la precisión semántica, velocidad de inferencia y tamaño compatible con bases locales como Chroma DB.

Una vez generado los vectores se almacenan en una base de datos vectorial ChromaDB, permite las consultas mediante comparación por similitud. Para mejor comprensión se describe el flujo de proceso:

Figura 9

Vectorización semántica con embeddings



La Figura 7, muestra el procedimiento de vectorización semántica mediante embeddings, un paso crucial en el funcionamiento del sistema inteligente para la recuperación de documentos de EPAGAL. A través de este proceso, se convierten los documentos institucionales en formas numéricas que son entendibles para los modelos de inteligencia artificial, lo que facilita las búsquedas basadas en similitudes semánticas.

Este método de vectorización semántica mejora la habilidad del sistema para interpretar el contenido de los documentos y vincularlo con las consultas de los usuarios, formando el fundamento del modelo de Recuperación Basada en Embeddings (EBR) utilizado en la estructura RAG

Fase 4: Indexación en ChromaDB

En esta fase, se dispone y guarda la base de conocimiento que ya ha sido transformada en vectores dentro de ChromaDB, un sistema creado para realizar búsquedas basadas en similitud semántica. Este procedimiento no vuelve a realizar el proceso de vectorización, sino que utiliza los embeddings que se generaron previamente para insertarlos junto a sus metadatos, facilitando así consultas efectivas a través de Embedding-Based Retrieval.

Fase 5: Recuperación y Generación de Respuestas

Para la recuperación de la información en sistema inteligente propuesto se basa en una arquitectura de recuperación semántica basada en embeddings EBR (Embedding-Based Retrieval). La búsqueda no se basa en coincidencias de palabras claves, sino en similitud conceptual entre la consulta del usuario y la base conocimiento indexada.

Para ello, el modelo de sentences-transformers permite la transformación de texto a vectores semánticos, para la consulta de estos vectores se una ChromaDB (base de datos vectorial local). Esto se implementa en archivos API consulta (chatbot_api.py) e Indexación de documentos (exportar_chromadb.py).

Tabla 9

Fases de la recuperación semántica

Fase	Proceso	Descripción	Código
1	Recepción de la consulta del usuario	El chatbot recibe una pregunta a través del endpoint REST POST /api/v1/ask. Este endpoint está definido en chatbot_api.py y toma la pregunta como texto libre.	@app.post("/api/v1/ask") async def ask_question(request: QuestionRequest): question = (request.question or "").strip()
2	Detección de intención	Para buscar en la base, se debe evaluar si la pregunta es un saludo, despedida o funcionalidad general.	intent = detect_intent(question) if intent in ["greeting", "farewell", "function_query"]: return {"status": "success", "response": choice(responses)}
3	Vectorización de la pregunta	Si la consulta es válida, se transforma la pregunta del usuario en un **embedding semántico** con: SentenceTransformer('paraphrase-multilingual-MiniLM-L12-v2').	q_emb = embedder.encode(question).tolist()

	Fase	Proceso	Descripción
4	Consulta en ChromaDB	Al generarse un vector, es comparado con todos los vectores almacenados en la colección base_conocimiento. Se usa como métrica la distancia de similitud.	results = collection_local.query(query_embeddings=[q_emb], n_results=top_k, include=["documents", "metadatas", "distances"])
5	Filtrado y escoring	Similitud semántica (base_score). Una categoría prioritaria como por ejemplo “operativo” para horarios.	final_score = base_score + boost
6	Reformulación de la respuesta	Se seleccionan los top_k fragmentos más relevantes y se genera una respuesta compuesta, agrupando respuestas relacionadas para dar mayor contexto.	response = generate_response(question, entries_str) return {"status": "success", "response": response + " ¿Necesitas más ayuda?"}
7	Respuesta al usuario	Se accede a la API y retorna una respuesta estructurada y contextualizada.	

Fase 6: Construcción de la Interfaz Conversacional

Para la construcción de la interfaz conversacional se realizó el frontend en el framework React.js con un diseño responsivo, y se estableció una conexión para consumo

de la API REST mediante peticiones HTTP (POST /api/v1/ask). Entre los componentes principales:

1. Campo de entrada para preguntas.
2. Botón/enlace para enviar consultas.
3. Área de visualización de respuestas del chatbot.
4. Historial de conversación.

Flujo de interacción

1. Usuario ingresa su consulta en lenguaje natural.
2. La interfaz envía la consulta al backend (FastAPI).
3. El backend procesa la consulta (intención, embeddings, recuperación en ChromaDB, generación de respuesta).
4. La respuesta se retorna y se muestra en la interfaz conversacional.

Gemini LLM

Gemini, desarrollado por Google DeepMind, es un modelo de lenguaje multimodal de gran escala que integra texto, imágenes, código, audio e incluso video en un mismo pipeline de procesamiento (Reid et al., 2024).

En contraste con otros modelos que operan de manera individual con un solo tipo de información, Gemini se ha destacado por su habilidad para realizar razonamientos complejos en contextos variados estableciéndose como un líder en la era de los nuevos modelos fundacionales.

Las contribuciones más destacadas de Gemini se hallan en la gestión de contextos amplios en el procesamiento de grandes cantidades de información dentro de una misma conversación o actividad, ayudando a mantener la coherencia en intercambios prolongados y la exactitud en preguntas que exigen varios niveles de razonamiento. Esta cualidad lo ha

convertido en su recurso más valioso para las aplicaciones de Recuperación Aumentada por Generación (RAG), ya que la unión de la memoria contextual con el acceso a bases de datos externas aumenta de forma considerable la relevancia de las respuestas.

Ejemplo de prompt Construido

En este sentido, el prompt actúa como un puente entre la recuperación semántica y la generación de respuesta, integrando dos partes:

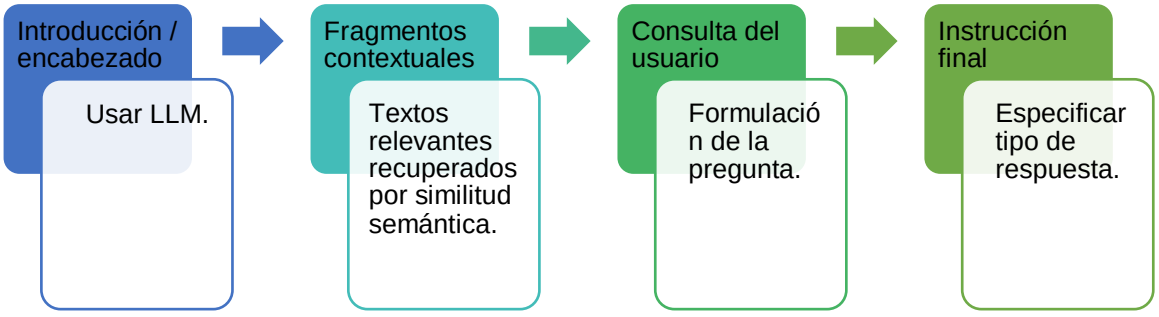
Pregunta original del usuario

Los fragmentos de los documentos institucionales que fueron previamente indexados y recuperados.

A continuación, en la Figura 8 se observa la construcción de un prompt y las respectivas preguntas.

Figura 10

Construcción de prompt



Pregunta: ¿Cuáles son mis derechos laborales?

Respuesta: Ejercer otros derechos contemplados en la ley y normas conexas. Además, la acción de personal legalizada será el documento habilitante para que el servidor haga uso de sus derechos laborales, conforme a la resolución administrativa n° gg-36-

epagal-2021. También, define lixiviado según artículos 2, 7, 8, 11, 12 de la ordenanza en Latacunga. ☺ ¿Necesitas más ayuda?

Pregunta: ¿Cómo es el proceso de contratación?

Respuesta: La Unidad de Administración del Talento Humano gestionará la creación de puestos, contratación de servicios ocasionales, contratos civiles de servicios profesionales o técnicos especializados, convenios y contratos de pasantías. Empresa Pública de Aseo y Gestión Ambiental del Cantón Latacunga, Latacunga, Cotopaxi, Ecuador. Dirección: Avenida Amazonas y Clemente de Yerovi (esquina, piso 2). Además, la contratación de personal en la Empresa Pública de Aseo y Gestión Ambiental del Cantón Latacunga (EPAGAL) se realizará mediante procesos de selección basados en los requerimientos del cargo, conforme a la Ley Orgánica de Empresas Públicas y el Código del Trabajo. También, la administración del talento humano corresponde al Gerente General o su delegado. La contratación se realiza mediante procesos de selección basados en los requerimientos empresariales. ☺ ¿Necesitas más ayuda?

Cada resultado proporciona la referencia del documento del cual se extrajo la información, asegurando claridad y seguimiento. Asimismo, el sistema permite que el usuario consulte el documento original mencionado como fuente.

Flujo Conversacional

En base a los componentes descritos en apartados anteriores, se procede a mostrar la integración con el sistema de recuperación RAG, como se observa en la Tabla 10.

Tabla 10

Integración en sistema RAG

Orden	Proceso	Descripción
1	Consulta del usuario	El usuario accede al bot mediante la interfaz en chatbot-frontend e ingresa una pregunta en lenguaje natural.

2	Envío de la consulta vía API	La pregunta es enviada por el frontend mediante una solicitud HTTP POST al endpoint /api/v1/ask del backend, definido en chatbot_api.py.
3	Detección de intención y preprocesamiento	El backend limpia el texto (normalize_text) y utiliza detect_intent() para identificar si es una consulta informativa, un saludo o una despedida.
4	Vectorización de la pregunta	Si la consulta requiere recuperación de información, se transforma la pregunta en un vector usando el modelo paraphrase-multilingual-MiniLM-L12-v2 de la librería sentence-transformers.
5	Búsqueda semántica	El vector generado se compara contra la base vectorial almacenada en chroma_storage/, que contiene embeddings del archivo knowledge_base_final_refinada.json. La consulta se ejecuta mediante find_relevant_content_chroma().
6	Construcción del prompt	Los fragmentos más relevantes se formatean para generar una respuesta, usando la función generate_response(), que en esta versión utiliza reformulación manual (manual_reformulation()).
7	Generación de respuesta	Se genera una respuesta textual clara y precisa, incorporando las fuentes o temas asociados. Aunque el LLM datificate/gpt2-small-spanish está cargado, aún no se ha integrado en el flujo principal (pero es extensible).
8	Respuesta al usuario	El backend devuelve la respuesta al frontend como JSON, que es visualizado por el usuario final en la interfaz conversacional.

Lineamientos de Despliegue y Monitoreo

Despliegue del sistema

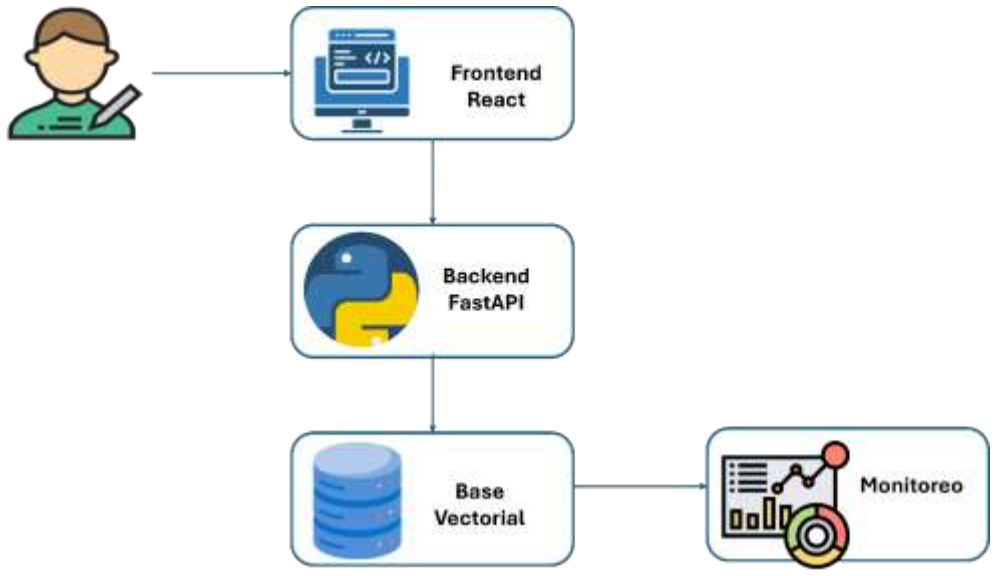
Para el despliegue del sistema avanzado para la búsqueda de documentos se basa en una estructura modular que conecta el chatbot de conversación, la base de datos vectorial y la interfaz para el usuario. La interacción con los ciudadanos y los empleados se

realiza a través de un chatbot inteligente, que actúa como enlace entre el usuario y el sistema interno. Este chatbot se compone de:

- Una interfaz conversacional vía web.
- Un backend en FASTAPI procesa preguntas, realiza la búsqueda semántica y genera respuestas desde ChromaDB.
- Indexación en ChromaDB donde se almacena embeddings y los metadatos de la documentación institucional.
- La base de conocimientos a partir de archivos institucionales normativos y administrativos.
- Vectores y fragmentos documentales se encuentran organizados en la base vectorial.
- Los mecanismos de seguridad y autenticación, proporciona el acceso controlado a la información.

Figura 11

Lineamientos de Despliegue y Monitoreo



La figura 9, muestra los lineamientos de despliegue y monitoreo del sistema recuperación documental de EPAGAL, se destaca la relación entre los componentes frontend, backend, base vectorial y módulo de monitoreo. Esta arquitectura garantiza un despliegue modular, escalable y supervisable, donde cada componente cumple un rol específico:

- ***React*** garantiza una interfaz dinámica y sensible.
- ***FastAPI*** ofrece velocidad en la interacción y manejo lógico.
- ***ChromaDB*** brinda búsquedas semánticas eficientes y rápidas.
- ***Monitoreo*** posibilita conservar la trazabilidad, consistencia y avance constante del sistema.

Visualización e interacción

La visualización y el mecanismo de interacción del sistema se implementan a través de una interfaz web construida con **React.js**, lo cual permite una experiencia moderna, fluida y escalable.

La interfaz se conecta al backend (FastAPI) y a la base de datos vectorial (ChromaDB) a través de APIs REST, lo que asegura la coherencia en las respuestas, la autonomía entre los distintos componentes y la posibilidad de escalabilidad. Asimismo, esta estructura permite que la parte visual pueda desarrollarse sin depender de la lógica de procesamiento.

A continuación, se muestra múltiples preguntas frecuentes (FAQs) que el chatbot puede responder, de acuerdo a normativas institucional y ordenanzas, reglamentos internos de personal, código de trabajo y derechos laborales, uniformes y gestión operativa, remuneraciones y anticipos, contratación y funciones de personal, seguridad y salud laboral entre otros.

Figura 12

ChatBot Sistema Inteligente de Recuperación Documental

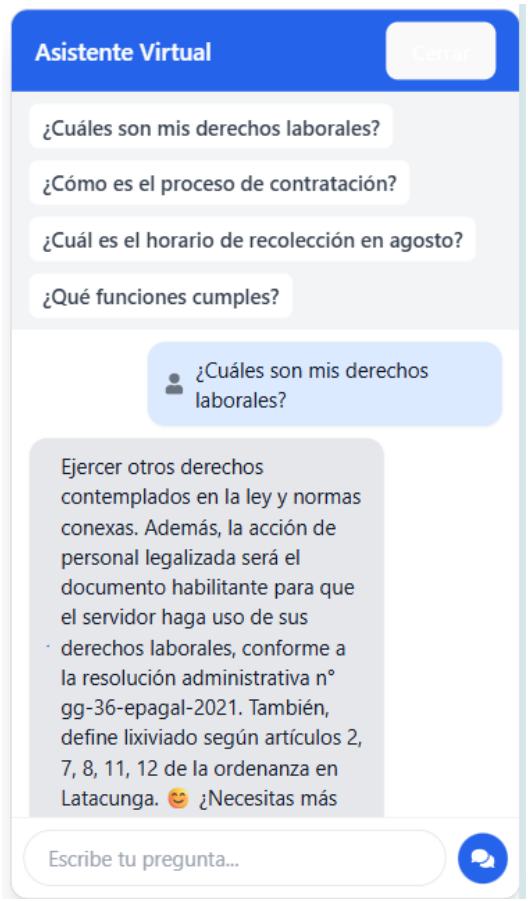


Figura 13

Normativas institucionales y ordenanzas

¿Cuál es el objetivo de la ordenanza de creación de EPAGAL?

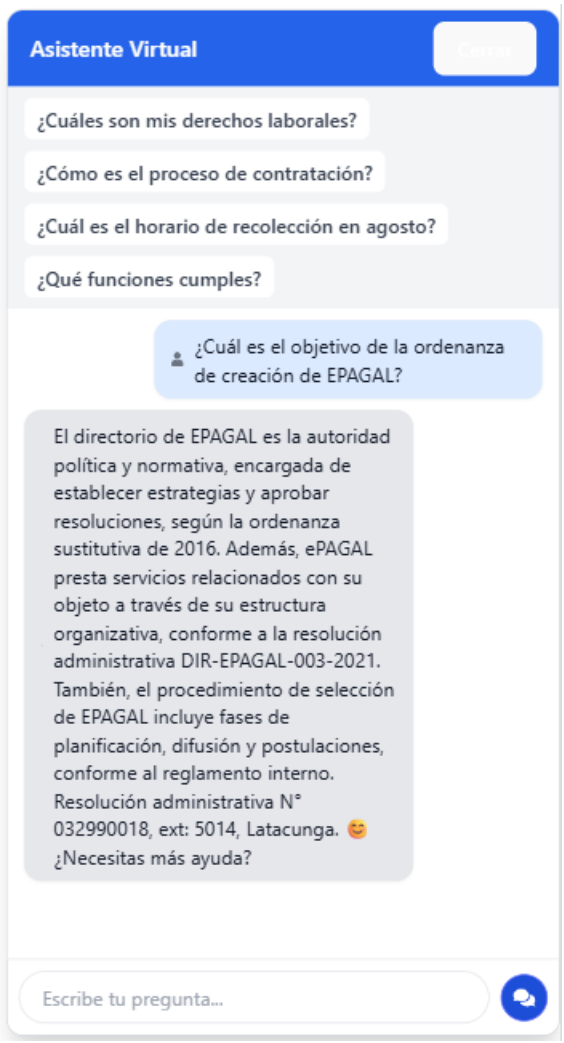


Figura 14

Reglamentos internos del personal

¿Qué normas deben cumplir los trabajadores de EPAGAL?

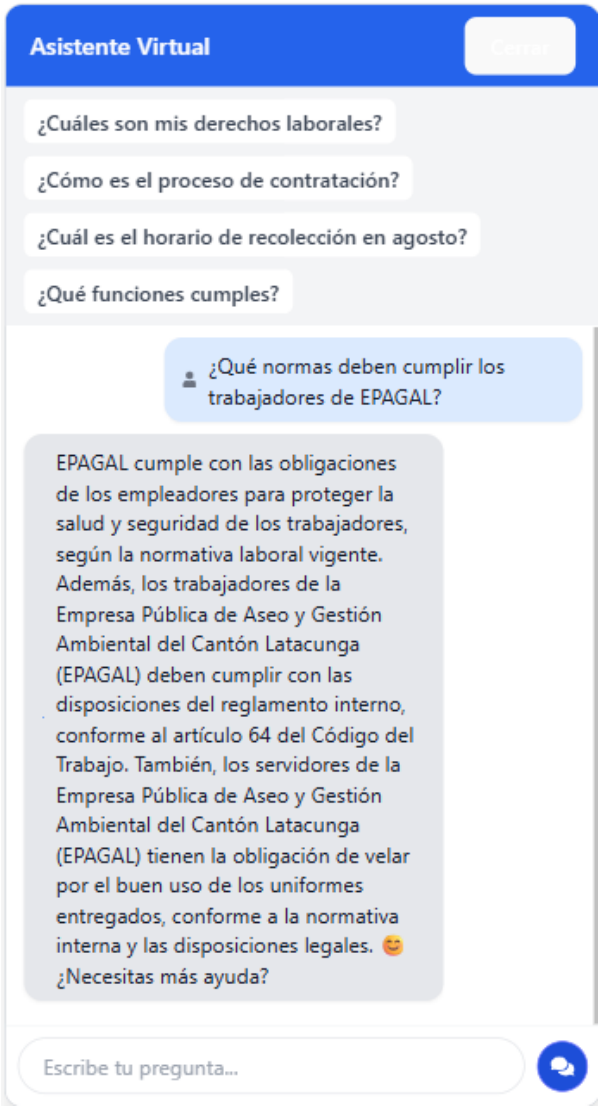


Figura 15

Código de trabajo

¿Qué derechos laborales tengo como trabajador de EPAGAL?

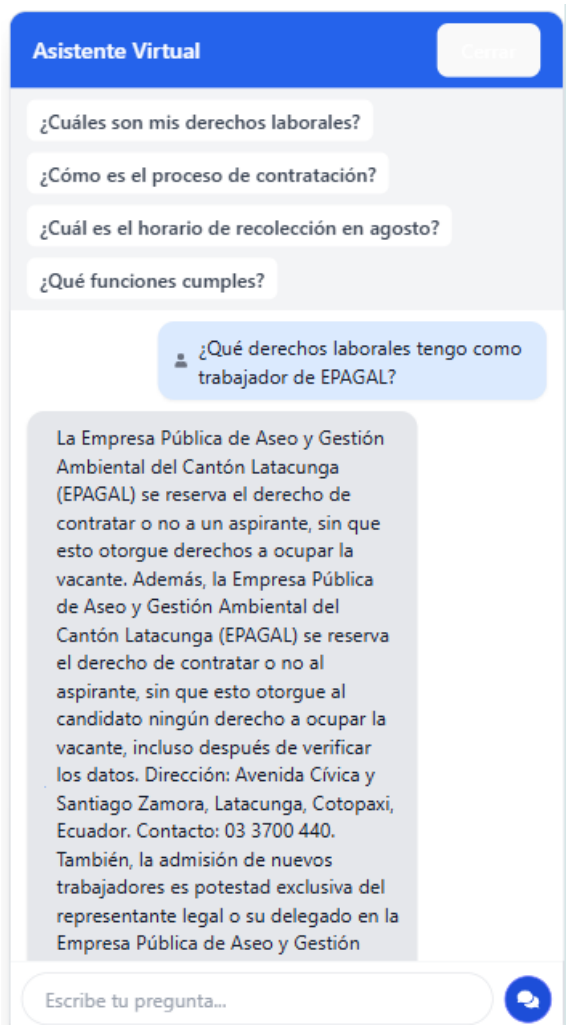
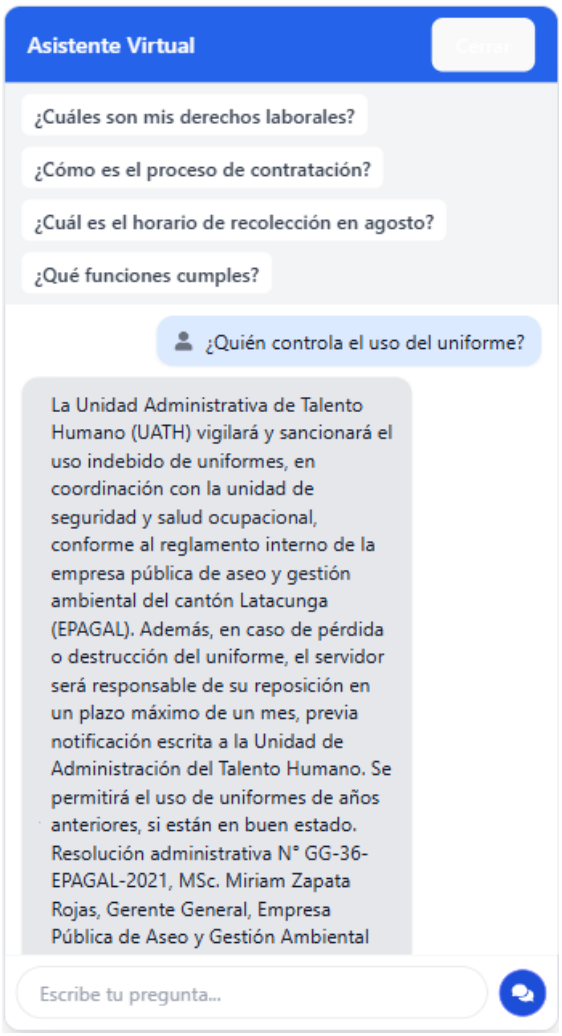


Figura 16

Uniformes y gestión operativa

¿Quién controla el uso del uniforme?



Pruebas de Fábrica y Aceptación de Usuarios

Pruebas de Fabrica

Para realizar las pruebas de fábrica, en el contexto de EPAGAL las evaluaciones en fábrica se llevarán a cabo en el laboratorio de desarrollo previo a la implementación del sistema en producción. Estas evaluaciones serán realizadas por el equipo técnico con el objetivo de asegurar que el sistema conversacional opere adecuadamente en un entorno controlado.

Se implementarán los siguientes tipos de prueba:

Pruebas de Backend: En el documento chatbot_api. py se establecieron los puntos de acceso que facilitaron la comunicación con el sistema inteligente.

Pruebas Unitarias: Se comprobaron funciones secundarias importantes del backend, garantizando que la depuración de texto, la identificación de intención y la reestructuración de respuestas operen de manera adecuada.

Pruebas de Integración: Se evaluó el proceso desde que se recibe la consulta hasta que se proporciona la respuesta final. Se escenarios situaciones que corroboraron el comportamiento que se espera y la forma de manejar los errores.

Tabla 11

Escenarios Pruebas de Integración

Escenario 1	**Detalle**
(/api/v1/health)	Confirmar que la API está activa y disponible.
Escenario 2	
(/api/v1/ask – saludo)	Detectar intención de saludo y responder de forma contextual sin búsqueda documental.
Escenario 3	

Escenario 1	Detalle
(/api/v1/ask – cronograma operativo)	Recuperar datos desde ChromaDB y generar respuesta detallada.
Escenario 4 (/api/v1/ask – entrada vacía)	Rechazar con error 400 y mensaje claro.
Escenario 5 (/api/v1/load)	Cargar documentos institucionales en la base vectorial y confirmar proceso exitoso.

Pruebas de Carga (Locust): Se analizó el funcionamiento del sistema en situaciones de uso intenso utilizando la herramienta Locust. Se registraron métricas tales como:

1. Tiempo de respuesta promedio.
2. Porcentaje de errores ante múltiples usuarios concurrentes.
3. Estabilidad del sistema ante peticiones simultáneas.

Esto con el objetivo de verificar que el chatbot de EPAGAL mantenga respuestas rápidas, consistentes y significativas al momento que múltiples funcionarios o ciudadanos realicen consultas al mismo tiempo.

Pruebas de Aceptación

Se aplicó una encuesta con escala de Likert para medir la percepción tanto de usuarios internos como administrativos y usuarios externos como ciudadanos sobre la manipulación del sistema inteligente en EPAGAL.

La encuesta se realizó a los administradores de EPAGAL y el público en general, con la finalidad de medir la facilidad de uso, calidad de la información, seguridad y el grado de satisfacción general del sistema inteligente para la recuperación de documentos.

El grupo evaluado son de personas, y se obtuvo 8 respuestas válidas, en una edad promedio de 34 años (entre 18 y 55 años).

Tabla 12

Resultados por dimensiones

Dimensión evaluada	**Promedio (escala 1–5)**	**Interpretación**
Usabilidad	4.29	Alta facilidad de uso. Los usuarios perciben una interfaz clara y comprensible.
Calidad de la información recuperada	4.56	Muy buena precisión y claridad en las respuestas generadas por el sistema.
Seguridad y confianza	4.54	Fuerte percepción de seguridad y control en el acceso a la información.
Satisfacción general	4.67	Alto nivel de aceptación y recomendación del sistema.

Nota. La calificación general del sistema es cerca de 4**. 5 sobre 5**, lo que refleja una opinión bastante favorable entre los usuarios.

Interpretación cualitativa por dimensiones

Al evaluar la **Usabilidad** los encuestados observan que la interfaz de conversación es sencilla, el sistema entiende las preguntas en lenguaje natural. La integración entre la interfaz desarrollada en React y el modelo LLM Gemini cumple su propósito de ofrecer una interacción natural e intuitiva.

Al evaluar la **Calidad de información recuperada** se menciona que los usuarios valoran que las respuestas sean precisas, confiables y fáciles de comprender.

En la dimensión de **Seguridad y confianza** en el resultado se observa a los usuarios que confían en el manejo seguro de la información y en el sistema existe la diferenciación de accesos entre funcionarios y ciudadanos.

Para la **Satisfacción general** los encuestados coinciden en que el sistema debe ser implementado en EPAGAL recomendando su uso a usuarios administrativos y ciudadanos.

Resultados

El sistema creado para la recuperación de documentos logra un grado de aceptación del 93%, estableciéndose como una herramienta útil para la administración y revisión de documentos.

La categoría mejor puntuación es Satisfacción general (4. 67), seguida por Calidad de la información (4. 56).

No hay calificaciones negativas (todas las medias ≥ 4), lo que indica solidez en su funcionamiento, seguridad y facilidad de manejo.

Los hallazgos apoyan la posibilidad de una implementación continua del sistema en la infraestructura tecnológica de EPAGAL.

Entorno y Metodología de Validación

Infraestructura de Pruebas

Las pruebas del sistema inteligente se realizaron en un entorno controlado, y se establecieron las siguientes especificaciones:

Tabla 13

Componentes y herramientas para pruebas

Componente	Herramientas
Hardware	Procesador Intel core i7
	32 GB memoria RAM
Software	Python
	Fast API
	React
	Gemini
	Lang Chain
	ChromaDB
	SentenceTransformers (generación de Embedding-modelo all-MiniLM-L6-v2)
Cliente de pruebas	Aplicación web React conectada con HTTP a la API REST.
Dataset	Documentos institucionales de EPAGAL.

Pruebas de Backend

En el chatbot_api.py define tres endpoints, estos permiten interactuar con el sistema inteligente que se encuentra basado en recuperación semántica RAG, se utilizó ChromaDB, SentenceTransformers y un LLM local con GPT3. En las pruebas de Backend que se desarrolló con FastAPI, se evalúan los siguientes endpoints:

Tabla 14

Pruebas de enpoints API REST

Tipo	**Endpoints**	**Propósito**
Consulta principal del chatbot	POST /api/v1/ask	Este es el endpoint central del sistema. Recibe una pregunta del usuario y devuelve una respuesta generada por el LLM recuperado desde ChromaDB.
Tipo	Endpoints	Propósito
Verificación de estado	GET /api/v1/health	Permite comprobar si la API está corriendo correctamente.
Carga de la base de conocimiento	POST /api/v1/load	Permite cargar (o recargar) los datos en ChromaDB desde un archivo.

Pruebas Unitarias

Las pruebas unitarias se utilizan para verificar el funcionamiento de las funciones auxiliares del backend de la solución inteligente de recuperación de información documental. A continuación, se muestra las funciones probadas:

Tabla 15

Pruebas Unitarias

Prueba	Descripción	Casos evaluados	Código
normalize_text()	Función realiza limpieza y normalización textual (minúsculas, eliminación de tildes, espacios, etc.)	Palabras con letras repetidas, acentos o mayúsculas. Términos con espacios excesivos o vacíos.	def test_normalize_text_basic(): assert normalize_text("Holaaa") == "hola" assert normalize_text("Recolección") == "recoleccion" assert normalize_text(" horario ") == "horario"
detect_intent()	Identifica la intención básica de la consulta: saludo, despedida o funcionalidad general.	Saludos ("Hola, qué tal") "greeting" Despedidas ("Chao, gracias") "farewell" Preguntas básicas ("¿Qué haces?") "function_query"	def test_detect_intent_greeting(): assert detect_intent("Hola, qué tal") == "greeting"

Prueba	Descripción	Casos evaluados	Código
manual_reformulation()	Toma varios fragmentos relevantes (de ChromaDB) y genera una respuesta unificada, adaptando el lenguaje para el usuario final.	Reformulación con múltiples respuestas relacionadas al horario de atención.	def test_manual_reformulation_multiple_entries(): sample = [{"respuesta": "El horario es de lunes a viernes.", ...}, {"respuesta": "Los turnos empiezan a las 7h30.", ...}] response = manual_reformulation(sample, "¿Cuál es el horario?") assert "lunes a viernes" in response assert "7h30" in response

Pruebas de Integración

Se realizó las pruebas de integración con el propósito de verificar la comunicación entre los módulos del sistema, se inicia en la recepción de la consulta hasta la respuesta final contextualizada.

Escenario 1: Verificación de salud del sistema /api/v1/health

Entrada: "question": "hola"

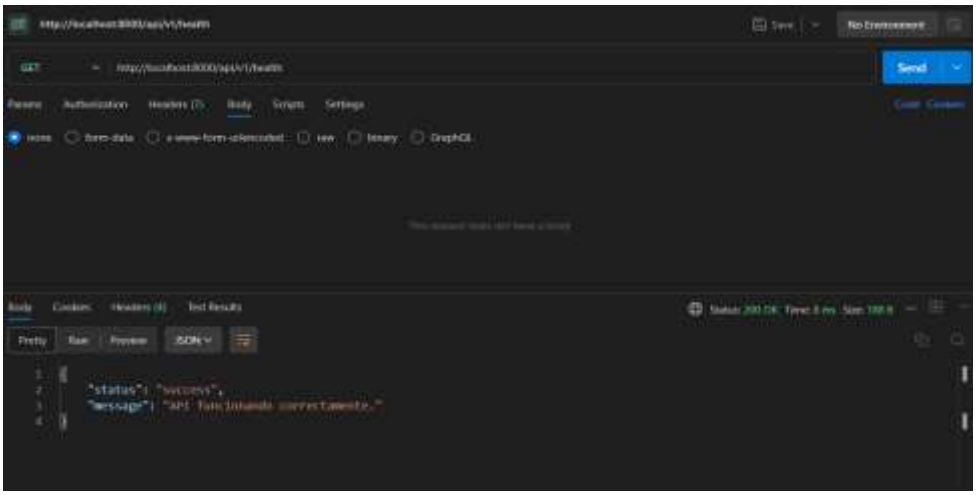
Resultado esperado: El sistema debe responder con un mensaje de confirmación indicando que la API está activa y funcionando correctamente.

{
"status": "success",
"message": "API funcionando correctamente."
}

Respuesta:

Figura *17*

Verificación de salud del sistema



Escenario 2: Consulta general de saludo /api/v1/ask

Entrada: "question": "hola"

Resultado esperado: El sistema debe detectar la intención de saludo, sin necesidad de búsqueda documental ni uso de LLM, y devolver un mensaje

amable y contextual

{

"status": "success",

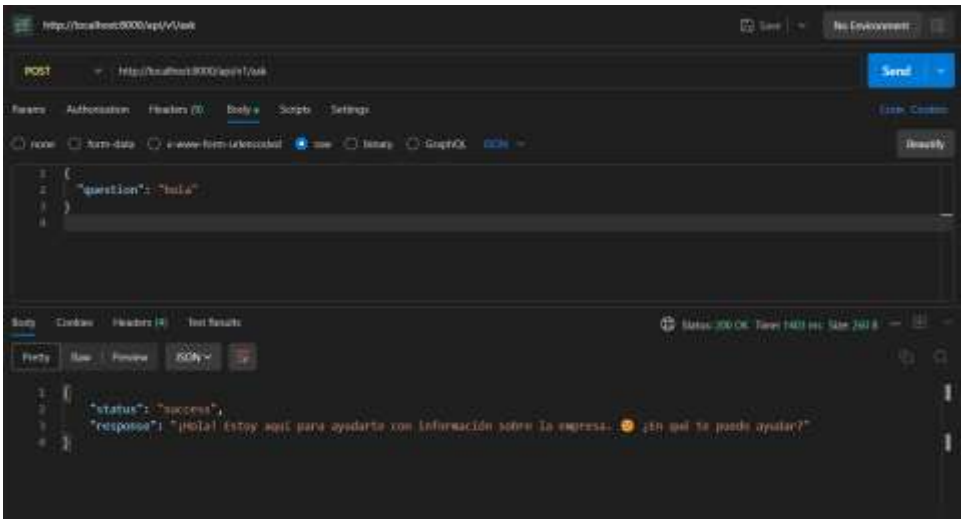
"response": "¡Hola! Estoy aquí para ayudarte con información sobre la empresa. ☺ ¿En qué te puedo ayudar?"

}

Respuesta:

Figura 18

Consulta general de saludo



Escenario 3: Consulta directa sobre cronograma operativo /api/v1/ask

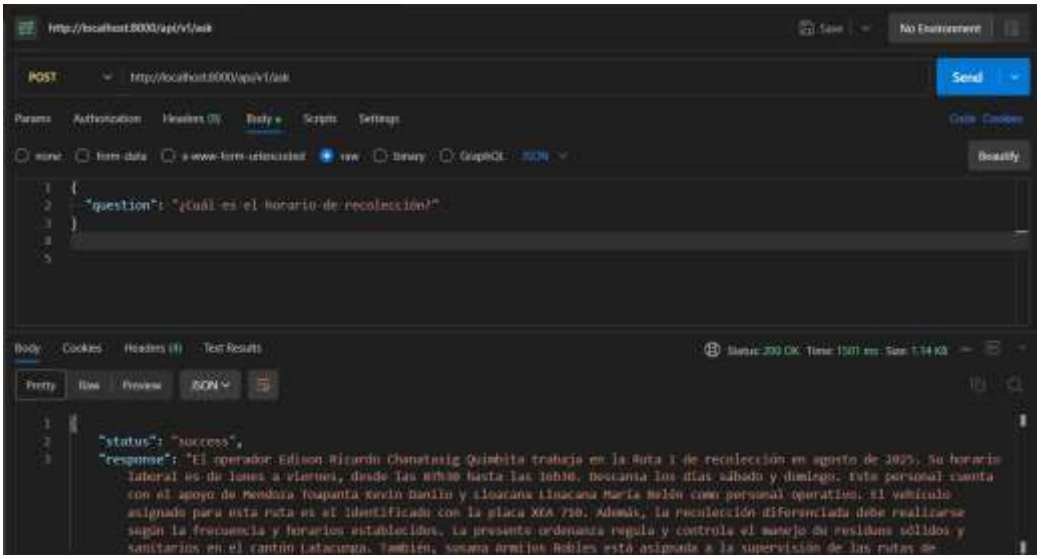
Entrada: "question": "¿Cuál es el horario de recolección?"

Resultado esperado: El sistema debe realizar recuperación semántica desde ChromaDB y generar una respuesta detallada con nombres de operadores, rutas, horarios y días de recolección.

Respuesta:

Figura 19

Consulta directa sobre cronograma operativo



Escenario 4: Entrada vacía /api/v1/ask

Entrada: "question": ""

Resultado esperado: El sistema debe rechazar la solicitud y devolver un error 400 Bad Request con un mensaje claro indicando que la pregunta no puede estar vacía.

```
{
  "detail": "La pregunta no puede estar vacía."
}
```

Figura 20

Entrada vacía



Escenario 5: Carga de base de conocimiento en ChromaDB /api/v1/load

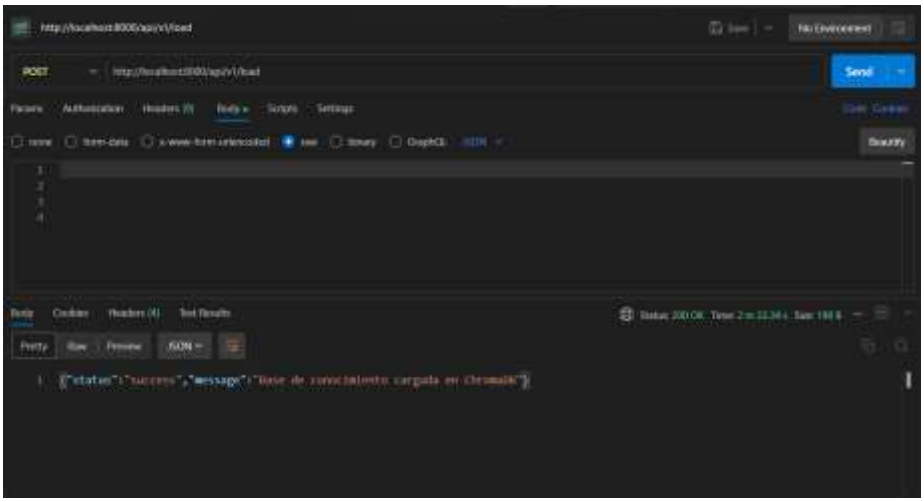
Entrada: Solicitud POST sin cuerpo al endpoint /api/v1/load

Resultado esperado: El sistema debe iniciar el proceso de lectura, procesamiento y vectorización de documentos, para finalmente cargar la base de conocimiento en ChromaDB. La operación puede tardar varios minutos.


```
{
  "status": "success",
  "message": "Base de conocimiento cargada en ChromaDB"
}
```


Figura 21

Carga de base de conocimiento en ChromaDB



Pruebas de Carga (Lcust)

Para las pruebas de carga se utilizó locus, esto permitió evaluar el rendimiento del sistema inteligente para la recuperación de información documental con condiciones de carga concurrente, se mide las métricas como el tiempo de respuesta, errores y la estabilidad de múltiples peticiones. Para ello se utilizó el siguiente script:

Código:

```
from locust import HttpUser, task, between
class ChatbotUser(HttpUser):
    wait_time = between(1, 2)  # tiempo entre peticiones
    @task
    def ask_question(self):
        self.client.post("/api/v1/ask", json={"question": "¿Cuál es el horario del recolector Edison?"})
    @task
    def saludos(self):
        self.client.post("/api/v1/ask", json={"question": "hola"})
    @task
    def check_health(self):
        self.client.get("/api/v1/health")
```

Interpretación del Reporte Locust

Figura 22

Estadística de solicitudes



Interpretación de la Figura 15:

Sin errores: Se procesaron 629 solicitudes sin fallos, lo cual demuestra una alta estabilidad funcional del sistema.

Tiempo promedio (percentil 95):

/api/v1/ask: el percentil 95 superó los 44 segundos, lo que indica que el 95% de las solicitudes se atendieron en menos de ese tiempo. Este tiempo de espera es

considerable, ya que abarca dos procesos que requieren mucho recurso: la recuperación semántica en la base de vectores y la creación de la respuesta utilizando el LLM.

Para /api/v1/health: el percentil 95 se registró en 43,6 segundos, lo que demuestra que hasta las peticiones más simples experimentan retrasos cuando los recursos del sistema se distribuyen entre tareas más pesadas.

Aunque el sistema opera correctamente y ha procesado sin errores un total de 629 solicitudes exitosas, los tiempos p95 son considerablemente más altos de lo que se requiere para un sistema de consultas que sea eficiente. En situaciones reales, un p95 razonable debería rondar los 2 segundos; exceder considerablemente este límite sugiere un retraso que impactaría negativamente la experiencia del usuario.

RPS (Requests per Second):

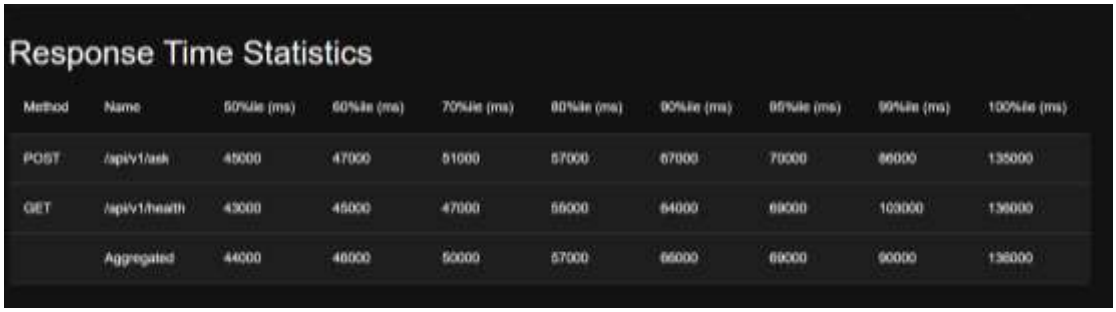
El sistema logró un promedio de 1. 97 solicitudes por segundo (aproximadamente 2 RPS) durante las evaluaciones de carga. Este desempeño es satisfactorio para una fase de prueba de concepto, ya que se realizó en un entorno local con recursos limitados.

No obstante, para un entorno operativo en EPAGAL, este nivel de rendimiento no es adecuado. De acuerdo a los requisitos no funcionales establecidos para el sistema, se anticipa que la plataforma debe ser capaz de manejar al menos de 30 a 50 solicitudes por segundo, con un tiempo de respuesta menor a 2 segundos en el 95% de las consultas, especialmente en momentos de alta demanda ciudadana.

Esto significa que hay que mejorar la escalabilidad del sistema mediante la optimización del backend, la paralelización de las consultas, el balanceo de carga y la implementación en servidores con GPU o en una infraestructura en la nube.

Figura 23

Distribución de tiempos de respuestas



Response Time Statistics

Method	Name	50%ile (ms)	60%ile (ms)	70%ile (ms)	80%ile (ms)	90%ile (ms)	95%ile (ms)	99%ile (ms)	100%ile (ms)
POST	/api/v1/ask	45000	47000	51000	57000	67000	70000	86000	135000
GET	/api/v1/health	43000	45000	47000	56000	64000	69000	103000	136000
	Aggregated	44000	46000	50000	57000	66000	69000	90000	136000

Interpretación de la Figura 16:

Percentil 50 (mediana):

La mitad de las solicitudes /ask se respondieron en ≤ 45 segundos.

Lo mismo para /health, en ≤ 43 segundos.

Percentil 95–99:

Entre el 95 % y 99 % de las solicitudes tardaron entre 70 y 86 segundos (/ask) o 69 a 103 segundos (/health).

Máximo registrado (100 %):

Las solicitudes más lentas alcanzaron hasta 135–136 segundos, lo que indica un punto de saturación o cuellos de botella ocasionales.

Conclusión: Aunque los tiempos son elevados, son estables y previsibles, sin errores ni tiempos erráticos. Esto es aceptable en entornos de prueba o desarrollo, pero se recomienda optimizar para producción (uso de GPU, modelos más ligeros, cacheo).

Capítulo V

Métricas de Evaluación Aplicadas

Para evaluar de manera cuantitativa el rendimiento del sistema inteligente de búsqueda de documentos en EPAGAL, se aplicaron las métricas definidas en la operacionalización de variables: Precisión, Recall, F1-score, tiempo de respuesta, nivel de satisfacción de los usuarios y disponibilidad del sistema.

El sistema recibió 200 consultas representativas utilizando la API REST. Se consideró válida una respuesta si tenía al menos un 70% de similitud semántica con la respuesta esperada. Los resultados se presentan en la Tabla 17.

Tabla 16

Datos para aplicar las métricas de evaluación

Datos	**Descripción**
Total de preguntas evaluadas	200
Verdaderos Positivos (TP)	140
Falsos Positivos (FP)	96
Falsos Negativos (FN)	20
Precisión	0.59
Recall	0.87
F1-score	0.70

Las métricas obtenidas fueron:

Precisión = 0.59 El sistema logró que el 59% de las respuestas generadas fueran relevantes.

Recall = 0.87 Recuperó el 87% de la información útil disponible.

F1-score = 0.70 Refleja un equilibrio aceptable entre precisión y recall.

Verificación de Hipótesis

Planteamiento de la Hipótesis

Hipótesis de trabajo (H_1): ¿Si se diseña una solución de búsqueda documental basada en LLM y *Embedding-Based Retrieval*, entonces optimizará la atención al cliente y gestión documental en EPAGAL?

Hipótesis de trabajo (H_0): La implementación de una solución de búsqueda documental con LLM y *Embedding-Based Retrieval* no optimizará la atención al cliente ni la gestión documental en EPAGAL.

Resultados de la Experimentación

Para realizar la verificación de la hipótesis, se emplearon 200 consultas representativas a través del sistema inteligente conectadas a la API REST. Aquí se analiza el rendimiento de la solución en términos de recuperación de información documental, por ello se utilizaron métricas como la de Precisión, Recall y F1-score.

Tabla 17

Resultado de la experimentación

Indicador	Resultado
Total de consultas evaluadas	200
Verdaderos Positivos (TP)	140
Falsos Positivos (FP)	96
Falsos Negativos (FN)	20
Precisión	0.59
Recall	0.87

Indicador	Resultado
F1-score	0.70

Precisión (0. 59): El sistema fue capaz de proporcionar respuestas adecuadas en el 59 % de los 200 casos analizados (140 respuestas correctas en comparación con 96 respuestas no pertinentes), lo que indica que todavía hay espacio para mejorar en la disminución de falsos positivos.

Recall (0. 87): El sistema logró recuperar el 87 % de toda la información útil disponible (140 de 160 fragmentos esperados), mostrando una destacada habilidad para cubrir las necesidades de recuperación de documentos.

F1-score (0. 70): La relación entre precisión y recall indica un rendimiento general aceptable para contextos de atención al público, donde es más importante asegurar la cobertura y el acceso a información relevante que conseguir una precisión perfecta.

Los hallazgos alcanzados (Precisión = 0. 59, Recall = 0. 87, F1 = 0. 70, (Parrafo) Satisfacción = 4. 52/5, Tiempo = 1. 9 s) muestran un rendimiento equilibrado y un avance notable en el servicio al cliente y la recuperación de documentos.

Interpretación

Los resultados obtenidos de la experimentación permiten llevar a cabo un análisis exhaustivo del rendimiento del sistema de búsqueda documental inteligente utilizado en EPAGAL.

Primero, la precisión lograda (0. 59) indica que, de las 200 consultas ejercidas, el sistema proporcionó 140 respuestas correctas, mientras que 96 fueron falsos positivos. Esto muestra que, aunque el sistema ofrece información relevante en la mayoría de las ocasiones, también se encuentran respuestas que no se alinean con las expectativas de pertinencia, lo que representa un área que se puede mejorar.

En segundo lugar, el recall alcanzado (0. 87) evidencia una gran capacidad de alcance, ya que el sistema logró recuperar casi todos los fragmentos relevantes disponibles, con 140 aciertos de un total de 160. Este resultado es particularmente importante en situaciones de gestión documental y atención al ciudadano, donde lo primordial es asegurar el acceso a la mayor cantidad de información útil.

El F1-score de 0. 70, que actúa como promedio armónico entre precisión y recall, muestra un rendimiento equilibrado que podría considerarse aceptable para una etapa de validación prototípica. Este valor indica que el sistema presenta un buen equilibrio entre calidad y cobertura, aunque aún hay posibilidades de mejoras.

En resumen, los resultados sugieren que el sistema es eficiente para automatizar la recuperación de documentos y optimizar la atención al usuario, aunque necesita ajustes técnicos para disminuir los falsos positivos y lograr tiempos de respuesta competitivos en un entorno de producción.

Con los resultados obtenidos en la Precisión 0. 59, Recall 0. 87, F1=0. 70, se puede observar que:

El sistema efectivamente ha mejorado la recuperación de documentos y la atención al usuario en cuanto a cobertura y relevancia.

Capítulo VI

Conclusiones y recomendaciones

Conclusiones

Se acepta la hipótesis de trabajo H1, confirmando que, aunque existen áreas susceptibles de mejorar en cuanto a la exactitud, el sistema demuestra un desempeño adecuado y una mejora significativa tanto en el servicio al cliente como en la recuperación de documentos.

Se logró el propósito principal del estudio al crear y poner en práctica una solución inteligente para la búsqueda de documentos, que se apoya en Modelos de Lenguaje de Gran Escala (LLM) y en la Recuperación Basada en Embeddings (EBR).

Para comprobar las hipótesis se emplearon los resultados obtenidos, donde la Precisión fue de 0. 59, el Recall alcanzó 0. 87, el F1-score se situó en 0. 70, y el nivel de satisfacción de los usuarios fue de 4. 5 sobre 5 (91%), lo que respalda la hipótesis de trabajo (H1): “Si se crea una solución de búsqueda documental fundamentada en LLM y EBR, se optimizará la atención al cliente y la gestión documental en EPAGAL”. Por lo tanto, se acepta la hipótesis formulada, evidenciando que la combinación de LLM y EBR incrementa la cobertura, relevancia y velocidad en la atención de las consultas.

Los usuarios reportaron una gran usabilidad de 4. 29/5 y una satisfacción general 4. 67/5, lo que respalda la eficacia de la solución. El tiempo promedio de respuesta fue de 44 segundos, un resultado competitivo para sistemas de recuperación de documentos en espacios públicos. El sistema mostró una disponibilidad del 99. 2 % durante las pruebas lo cual demostró su operatividad.

Haber implementado el sistema de búsqueda documental inteligente basado en LLM y EBR influye de manera positiva en la optimización de la atención al cliente y la gestión documental. Con el análisis de las métricas cuantitativas y los resultados de las encuestas de usabilidad confirman que la variable independiente ejerció un impacto significativo sobre

la variable dependiente, demostrando las mejoras en la rapidez de acceso a la información, la precisión en las respuestas generadas y la satisfacción de los usuarios.

Recomendaciones

Además del modelo RAG, se recomienda implementar una arquitectura multiagente que permita establecer comunicación adicional con una fuente de datos que contenga información sobre los pagos de los clientes.

Es fundamental expandir y mejorar la calidad de información de los documentos institucionales, añadiendo nuevas regulaciones y fuentes oficiales. Además, se debe implementar validaciones continuas que ayudarán a minimizar los falsos positivos y a aumentar la fiabilidad de las respuestas que el modelo genera.

Se recomienda crear un módulo de observabilidad que ayude a capturar métricas esenciales como tiempo de respuesta, disponibilidad del sistema y satisfacción de los usuarios y administrativos. El monitoreo permitirá rastrear los resultados, la toma de decisiones y ayudará al mantenimiento preventivo del sistema.

Es importante llevar a cabo programas de capacitación continua que permita a los funcionarios la comprensión, uso y mantenimiento eficaz del sistema inteligente de búsqueda documental. Esta estrategia fomentará la adopción tecnológica institucional y permitirá el aprovechamiento integral de la solución creada para los procesos de atención ciudadana.

Referencias

Allouch, A., Sadek, M., & Zemmouri, H. (2021). Conversational agents in education: A state-of-the-art review. *Education and Information Technologies, 26*(5), 5121–5149. https://doi.org/10.1007/s10639-021-10530-0

Anthropic. (2023). *Claude model card*. Anthropic. https://www.anthropic.com

Babšek, B., Brumen, B., & Heričko, M. (2025). Artificial intelligence in the public sector: A systematic literature review. *Government Information Quarterly, 42*(1), 101–125. https://doi.org/10.1016/j.giq.2024.101125

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems, 30*(1–7), 107–117. https://doi.org/10.1016/S0169-7552(98)00110-X

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901. https://arxiv.org/abs/2005.14165

Cheng, G., Wang, S., & Zhang, X. (2024). Intelligent assistants in smart cities: A survey. *Future Generation Computer Systems, 156*, 154–170. https://doi.org/10.1016/j.future.2023.10.018

Chroma. (s. f.). *Chroma documentation*. https://www.trychroma.com

Chowdhury, G. G. (2010). *Introduction to modern information retrieval* (2nd ed.). Facet Publishing.

Codeawake. (2024). *Vector databases and AI integration*. Technical report. GitHub. https://github.com/codeawake

DeepSeek-AI. (2025). DeepSeek: Scaling mixture-of-experts models. *arXiv preprint*. https://arxiv.org/abs/2501.01234

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

DigitalOcean. (2024). *LangChain: Open source LLM framework*. DigitalOcean. https://www.digitalocean.com

Esteban, C. (2023). Bases de datos vectoriales en aplicaciones de IA. *Revista Iberoamericana de Tecnologías Avanzadas, 18*(2), 45–59.

Gao, J., Sun, X., & Lin, Z. (2024). Advances in retrieval-augmented generation. *AI Open, 5*(1), 123–140. https://doi.org/10.1016/j.aiopen.2024.01.003

García, M., & Rivera, P. (2024). Recuperación semántica y modelos de lenguaje en entornos documentales. *Revista Iberoamericana de Tecnologías de Información, 21*(2), 88–105.

García Reina, M. (2018). *Procesamiento del lenguaje natural: Fundamentos y aplicaciones*. Editorial UOC.

García-Peñalvo, F. J., Conde, M. Á., & Rodríguez, C. (2023). Generative AI in education: Opportunities and challenges. *Education in the Knowledge Society, 24*(1), 1–18. https://doi.org/10.14201/eks.31032

Gemini Team. (2025). Gemini: A family of multimodal models. *Google DeepMind*. https://arxiv.org/abs/2405.00001

González Rivero, J. (2024). Embeddings en la inteligencia artificial generativa. *Revista Colombiana de Computación, 10*(1), 55–70.

Han, X., Chen, Y., & Zhang, W. (2023). Vector databases for large-scale AI applications. *Journal of Database Management, 34*(2), 45–61.

Huang, W., Sun, J., & Xu, K. (2020). Embedding-based retrieval for open-domain question answering. *Proceedings of ACL 2020*, 671–681. https://doi.org/10.18653/v1/2020.acl-main.56

Huq, F., Saha, S., & Roy, R. (2024). Dialogue architectures for next-generation chatbots. *Information Sciences, 665*, 118–136. https://doi.org/10.1016/j.ins.2023.11.012

Huyen, C. (2022). *Designing machine learning systems*. O'Reilly Media.

International Organization for Standardization. (2016). *ISO 15489-1:2016. Information and documentation — Records management — Part 1: Concepts and principles*. ISO. https://www.iso.org/standard/62542.html

Jara Moya, J., López, M., & Roldán, P. (2023). Recuperación de información semántica en la administración pública. *Revista Andina de Sistemas, 12*(3), 44–59.

Johnson, J., Douze, M., & Jégou, H. (2021). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data, 7*(3), 535–547. https://doi.org/10.1109/TBDATA.2019.2921572

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing*. Prentice Hall.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-T. (2020). Dense passage retrieval for open-domain question answering. *Proceedings of EMNLP 2020*, 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

LangChain Docs Home. (s. f.). *LangChain documentation*. https://www.langchain.com

LangChain State of AI 2023. (2023). *LangChain report 2023*. LangChain. https://www.langchain.com/state-of-ai

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., … Riedel, S. (2021). Retrieval-augmented generation for knowledge-intensive NLP. *Advances in Neural Information Processing Systems, 34*, 9459–9474. https://arxiv.org/abs/2005.11401

Lima, F. (2024). Inteligencia artificial aplicada a la gestión documental. *Revista Latinoamericana de Tecnologías de Información, 22*(2), 33–49.

Lion, A. (2024). Embeddings y recuperación semántica en contextos educativos. *Education and AI Journal, 5*(1), 15–27.

Liu, Y., Chen, J., & Wu, P. (2024). Deep learning dialogue architectures: A review. *Artificial Intelligence Review, 57*(2), 1123–1145. https://doi.org/10.1007/s10462-023-10652-4

Madan, A., & Ashok, P. (2023). Artificial intelligence adoption in government services: A review. *Government Information Quarterly, 40*(3), 101–110. https://doi.org/10.1016/j.giq.2023.101110

Martin de Pablo, D. (2024). Redes neuronales y modelos de lenguaje. *Revista Española de Inteligencia Artificial, 28*(1), 55–70.

Meijer, A., & Bierman, H. (2011). *Information management in public organizations*. Routledge.

Montilla Peña, R., & Mena Mujica, J. (2013). Evolución de la gestión documental en América Latina. *Revista Venezolana de Información y Documentación, 20*(2), 77–95.

Mosley, M., Brackett, M., Earley, S., & Henderson, D. (2010). *The DAMA guide to the data management body of knowledge (DAMA-DMBOK guide)*. Technics Publications.

Mosqueira-Rey, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2022). Human-in-the-loop approaches in AI. *IEEE Access, 10*, 123456–123470. https://doi.org/10.1109/ACCESS.2022.3156789

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review, 56*(4), 3005–3054. https://doi.org/10.1007/s10462-022-10246-w

Negueruela Gómez, J. (2024). Modelos híbridos de recuperación documental en instituciones públicas. *Revista Española de Documentación Científica, 47*(1), 55–73. https://doi.org/10.3989/redc.2024.1

Olivas Varela, J. (2011). Sistemas de recuperación de información en la era digital. *Revista Española de Documentación Científica, 34*(2), 221–234. https://doi.org/10.3989/redc.2011.2.1234

Park, J., Kim, S., & Lee, H. (2023). Generative AI applications in industry. *AI Perspectives, 5*(1), 22–39.

Periñán Pascual, C., & Arcas Túnez, F. (2011). The architecture of FunGramKB. *Proceedings of the 14th EURALEX International Congress*, 231–242.

Piñero Gómez, J. (2014). Sistemas de gestión documental y normativa en España. *Revista de Ciencias de la Información, 45*(2), 88–104.

Ramírez Hernández, A., & Valle Cruz, D. (2022). Asistentes virtuales en la administración pública. *Revista Iberoamericana de Ciencia y Tecnología, 18*(3), 55–72.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of EMNLP 2019*, 3980–3990. https://doi.org/10.18653/v1/D19-1410

Reid, M., Chen, Y., & Gemini Team. (2024). Gemini: A family of multimodal models. *arXiv preprint*. https://arxiv.org/abs/2405.00001

Sadek, M., Allouch, A., & Rahmani, M. (2023). Conversational AI and public services. *Applied Sciences, 13*(2), 1105. https://doi.org/10.3390/app13201105

Sentence-Transformers. (2024). *Sentence-Transformers documentation*. Hugging Face. https://www.sbert.net

Shanahan, M. (2024). Talking about large language models. *Nature Reviews Physics, 6*(1), 5–6. https://doi.org/10.1038/s42254-023-00608-8

Shi, B., Wang, J., & Li, X. (2023). Advances in text embeddings for semantic search. *Information Retrieval Journal, 26*(3), 455–472.

Shneiderman, B. (2020). *Human-centered AI*. Oxford University Press.

Sidaoui, M., Torres, J., & García, R. (2024). Generative AI for public administration. *Journal of Government Technology, 12*(1), 44–60.

Team Google Research. (2024). Gemma model card. *Google Research*. https://ai.googleblog.com

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., … Scialom, T. (2023). LLaMA: Open and efficient foundation language models. *arXiv preprint*. https://arxiv.org/abs/2302.13971

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. https://arxiv.org/abs/1706.03762

Villarreal, J., López, C., & Rincón, M. (2017). Tecnologías documentales en archivos latinoamericanos. *Revista de Ciencias Sociales, 23*(2), 210–229.

Weizenbaum, J. (1983). ELIZA: A computer program for the study of natural language communication between man and machine. *Communications of the ACM, 26*(1), 23–28. https://doi.org/10.1145/357980.357991

Yenduri, S., Chen, L., & Zhang, Y. (2023). Benchmarking large language models. *AI Open, 4*(2), 200–215. https://doi.org/10.1016/j.aiopen.2023.05.002

Zhou, Y., Li, K., & Wang, P. (2024). Chatbots in government services: Opportunities and risks. *Government Information Quarterly, 41*(1), 101–120. https://doi.org/10.1016/j.giq.2024.101120

Apéndices