UNIVERSIDAD DE LAS FUERZAS ARMADAS - ESPE

Maestría en Ingeniería en Software

Tema:

“Diseño de una solución de búsqueda de información documental con modelos LLM y Embedding-Based Retrieval, para atención al cliente en EPAGAL”

Autor:

Salgado Quimbita, Juan Mauricio

Director: Gallardo Corrales, Diego Eduardo. Msc.

LATACUNGA, AGOSTO 2026



Agenda

- **Introducción**
- **Problemática**
- **Objetivos**
- **Justificación e importancia**
- **Alcance**
- **Marco teórico**
- **Propuesta metodológica**
- **Arquitectura del sistema**
- **Desarrollo**
- **Implementación**
- **Resultados y métricas**
- **Conclusiones y recomendaciones**

Introducción

- EPAGAL enfrenta dificultades para acceder de manera rápida a normativas y documentos internos.
- Se propone un sistema inteligente basado en LLM y EBR para optimizar la gestión documental.
- La arquitectura utiliza el enfoque RAG (Retrieval-Augmented Generation) combinando IA generativa y recuperación semántica.

- **Introducción**
- Problemática
- Objetivos
- Justificación e importancia
- Alcance
- Marco teórico
- Propuesta metodológica
- Arquitectura del sistema
- Desarrollo
- Implementación
- Resultados y métricas
- Conclusiones y recomendaciones

Problemática

- Los sistemas actuales son manuales y no permiten búsquedas ágiles ni contextualizadas.
- Dificultad en localizar reglamentos, ordenanzas y procesos administrativos.
- Se requiere una solución que integre inteligencia artificial y recuperación documental para mejorar la atención al cliente.

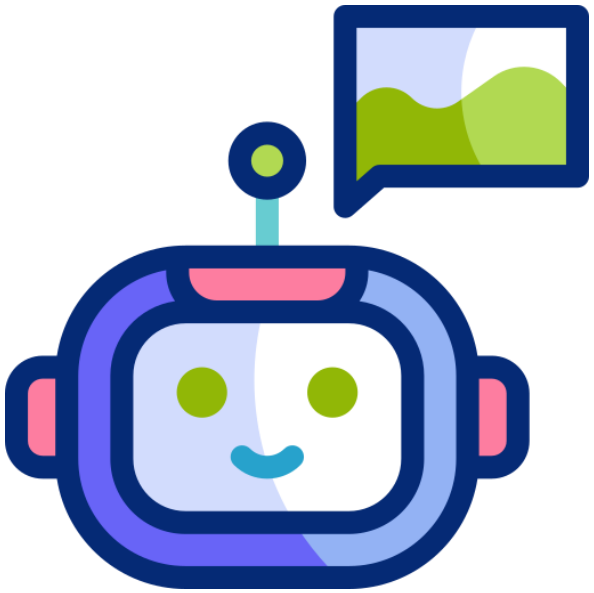
Pregunta de investigación:

¿Cómo mejorar la atención al cliente en EPAGAL mediante el uso de Inteligencia Artificial optimizando la búsqueda y recuperación de información?

- Introducción
- **Problemática**
- Objetivos
- Justificación e importancia
- Alcance
- Marco teórico
- Propuesta metodológica
- Arquitectura del sistema
- Desarrollo
- Implementación
- Resultados y métricas
- Conclusiones y recomendaciones

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.



- Introducción
- Problemática
- **Objetivos**
- Justificación e importancia
- Alcance
- Marco teórico
- Propuesta metodológica
- Arquitectura del sistema
- Desarrollo
- Implementación
- Resultados y métricas
- Conclusiones y recomendaciones

Objetivos específicos

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

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

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

- Introducción
- Problemática
- **Objetivos**
- Justificación e importancia
- Alcance
- Marco teórico
- Propuesta metodológica
- Arquitectura del sistema
- Desarrollo
- Implementación
- Resultados y métricas
- Conclusiones y recomendaciones

Justificación e importancia

- EPAGAL maneja grandes volúmenes de documentos sin un sistema de búsqueda eficiente.
- Los LLM y EBR permiten consultas en lenguaje natural, ofreciendo respuestas precisas.
- Impacto: mejora la eficiencia institucional, la transparencia y la satisfacción ciudadana.

- Introducción
- Problemática
- Objetivos
- **Justificación e importancia**
- Alcance
- Marco teórico
- Propuesta metodológica
- Arquitectura del sistema
- Desarrollo
- Implementación
- Resultados y métricas
- Conclusiones y recomendaciones

Alcance



Aspectos claves:

1. Implementación de una interfaz conversacional con búsqueda semántica.
2. Base de conocimiento: reglamentos, ordenanzas y procesos de EPAGAL.
3. Evaluación de desempeño mediante métricas de precisión, recall y satisfacción.
4. No incluye automatización total de procesos administrativos.

- Introducción
- Problemática
- Objetivos
- Justificación e importancia
- **Alcance**
- Marco teórico
- Propuesta metodológica
- Arquitectura del sistema
- Desarrollo
- Implementación
- Resultados y métricas
- Conclusiones y recomendaciones

Marco Teórico

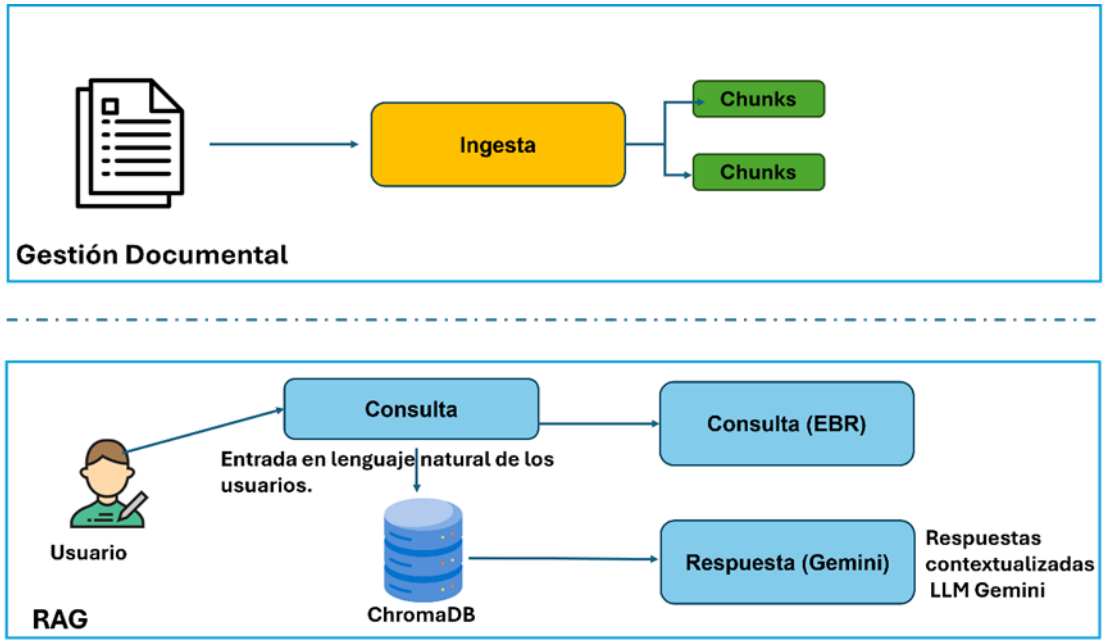
- Gestión Documental basada en el marco DAMA-DMBOK.
- Evolución: archivos físicos → bases relacionales → web semántica → LLM + RAG.
- Tecnologías base:
  - Python
  - FastAPI
  - LangChain
  - ChromaDB
  - Gemini LLM



- Introducción
- Problemática
- Objetivos
- Justificación e importancia
- Alcance
- **Marco teórico**
- Propuesta metodológica
- Arquitectura del sistema
- Desarrollo
- Implementación
- Resultados y métricas
- Conclusiones y recomendaciones

Propuesta metodológica

Figura: Integración entre gestión documental y RAG.

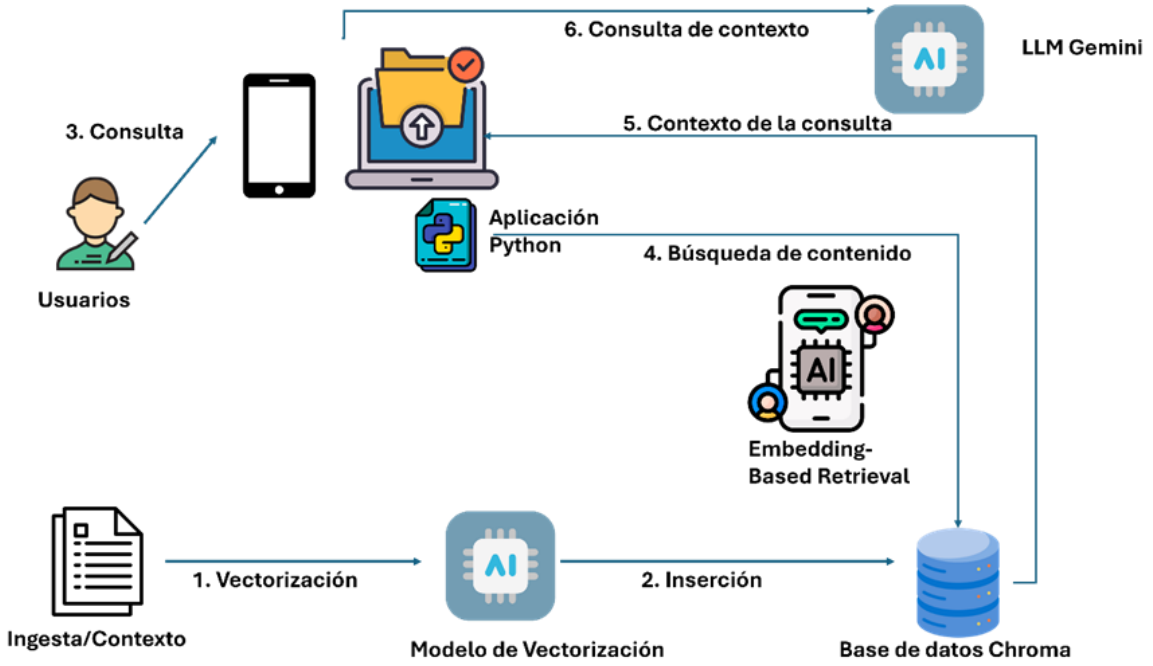


- Integración de RAG con la gestión documental de EPAGAL.
- Documentos institucionales se transforman en embeddings y se almacenan en ChromaDB.
- Consultas en lenguaje natural → recuperación semántica → respuesta generada por Gemini LLM.

- Introducción
- Problemática
- Objetivos
- Justificación e importancia
- Alcance
- Marco teórico
- **Propuesta metodológica**
- Arquitectura del sistema
- Desarrollo
- Implementación
- Resultados y métricas
- Conclusiones y recomendaciones

Arquitectura del sistema

Figura: Arquitectura Sistema Inteligente



Visualiza la arquitectura lógica del sistema RAG aplicado a la gestión documental institucional, mostrando cómo se integran tecnologías de IA (LangChain, ChromaDB, Gemini) para ofrecer un servicio de búsqueda y atención inteligente.

- Introducción
- Problemática
- Objetivos
- Justificación e importancia
- Alcance
- Marco teórico
- Propuesta metodológica
- **Arquitectura del sistema**
- Desarrollo
- Implementación
- Resultados y métricas
- Conclusiones y recomendaciones

Desarrollo

- Introducción
- Problemática
- Objetivos
- Justificación e importancia
- Alcance
- Marco teórico
- Propuesta metodológica
- Arquitectura del sistema
- **Desarrollo**
- Implementación
- Resultados y métricas
- Conclusiones y recomendaciones

Implementación

Fase 1: Preprocesamiento y vectorización.

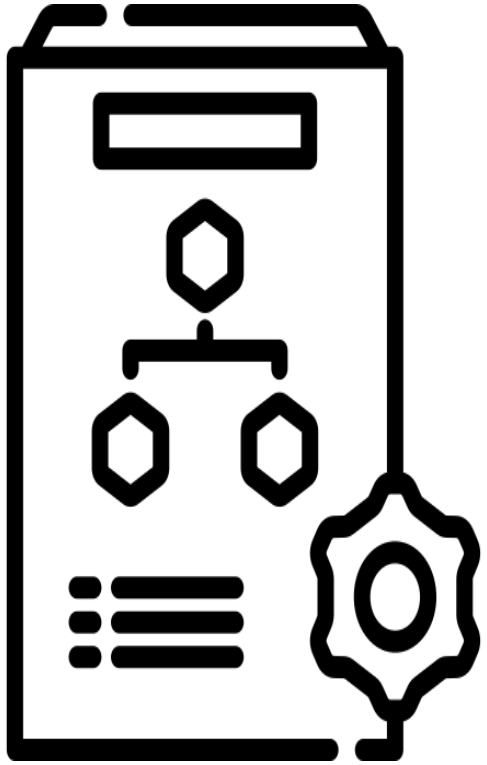
Fase 2: Limpieza y segmentación de fragmentos.

Fase 3: Vectorización semántica con Sentence-Transformers.

Fase 4: Indexación en ChromaDB.

Fase 5: Recuperación y generación de respuestas (API REST).

Fase 6: Interfaz conversacional (React + FastAPI).



- Introducción
- Problemática
- Objetivos
- Justificación e importancia
- Alcance
- Marco teórico
- Propuesta metodológica
- Arquitectura del sistema
- Desarrollo
- **Implementación**
- Resultados y métricas
- Conclusiones y recomendaciones

Resultados y métricas

Contexto de validación

El sistema fue evaluado en un entorno controlado de pruebas en EPAGAL, considerando dos dimensiones complementarias:

Validación técnica (métricas de rendimiento):

Medición de la precisión, recall y F1-score mediante consultas automatizadas a la API REST.

Validación de usuario (métricas de aceptación): Evaluación de usabilidad, satisfacción y percepción de seguridad mediante una encuesta con escala Likert aplicada a 8 usuarios (4 administrativos y 4 ciudadanos).

- Introducción
- Problemática
- Objetivos
- Justificación e importancia
- Alcance
- Marco teórico
- Propuesta metodológica
- Arquitectura del sistema
- Desarrollo
- Implementación
- **Resultados y métricas**
- Conclusiones y recomendaciones

Resultados y métricas

Análisis cuantitativo

Indicador	Descripción	Resultado	Interpretación
Número de consultas evaluadas	Total de preguntas procesadas por el sistema durante las pruebas REST.	**200**	Muestra un volumen representativo de interacciones para validar consistencia.
Precisión (0.59)	Proporción de respuestas correctas respecto al total de respuestas generadas.	**59 %**	Indica que el sistema proporciona respuestas correctas en la mayoría de los casos, aunque se presentan falsos positivos (respuestas parcialmente relevantes).
Recall (0.87)	Proporción de información relevante recuperada respecto al total disponible en la base documental.	**87 %**	Demuestra una alta capacidad del sistema para cubrir casi toda la información útil, lo cual es prioritario en contextos públicos.
F1-Score (0.70)	Promedio armónico entre precisión y recall.	**0.70**	Representa un equilibrio adecuado entre exactitud y cobertura, válido para entornos institucionales.
Satisfacción del usuario (4.67 / 5)	Promedio obtenido en encuestas de usabilidad (escala Likert 1–5).	**93 % de aceptación**	Evidencia que los usuarios perciben la plataforma como clara, confiable y funcional.
Disponibilidad del sistema (99.2 %)	Tiempo operativo sin fallos durante las pruebas de carga y monitoreo.	**Alta estabilidad**	Refleja solidez en la arquitectura modular RAG y la integración FastAPI–ChromaDB–Gemini.

- Introducción
- Problemática
- Objetivos
- Justificación e importancia
- Alcance
- Marco teórico
- Propuesta metodológica
- Arquitectura del sistema
- Desarrollo
- Implementación
- **Resultados y métricas**
- Conclusiones y recomendaciones

Conclusiones

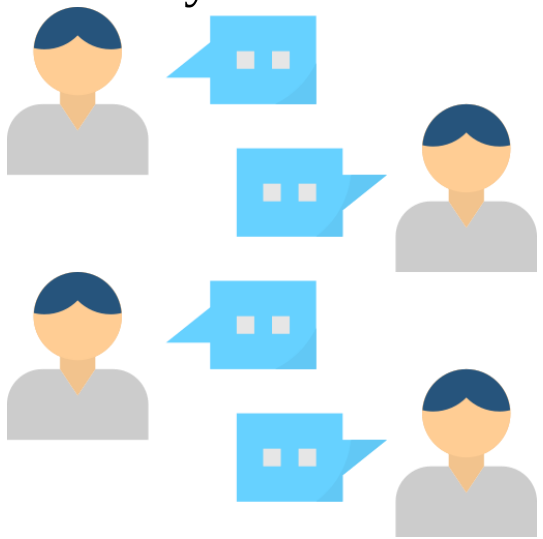
- Se confirma la hipótesis: el sistema mejora la atención al cliente y la recuperación documental.
- La combinación LLM + EBR permite respuestas contextuales y precisas.
- La arquitectura RAG demostró ser eficiente y escalable para EPAGAL.



- Introducción
- Problemática
- Objetivos
- Justificación e importancia
- Alcance
- Marco teórico
- Propuesta metodológica
- Arquitectura del sistema
- Desarrollo
- Implementación
- Resultados y métricas
- **Conclusiones y recomendaciones**

Recomendaciones

- Incorporar arquitectura multiagente para conectar con fuentes de pagos y clientes.
- Mejorar calidad documental y validaciones automáticas.
- Implementar módulo de monitoreo y métricas.
- Capacitar al personal para el uso y mantenimiento del sistema.



- Introducción
- Problemática
- Objetivos
- Justificación e importancia
- Alcance
- Marco teórico
- Propuesta metodológica
- Arquitectura del sistema
- Desarrollo
- Implementación
- Resultados y métricas
- **Conclusiones y recomendaciones**

Gracias por su atención

juan.salgado@espe.edu.ec