Ciencia y Tecnología
Secretaría de Ciencia, Humanidades, Tecnología e Innovación

Transformación Digital
Agencia de Transformación Digital y Telecomunicaciones

Foro Interinstitucional de Inteligencia Artificial y Supercómputo

11 y 12 de junio de 2025

2025
Año de
La Mujer
Indígena

Ciencia y Tecnología
Secretaría de Ciencia, Humanidades, Tecnología e Innovación

Transformación Digital
Agencia de Transformación Digital y Telecomunicaciones

Asistente para Tramites y Servicios

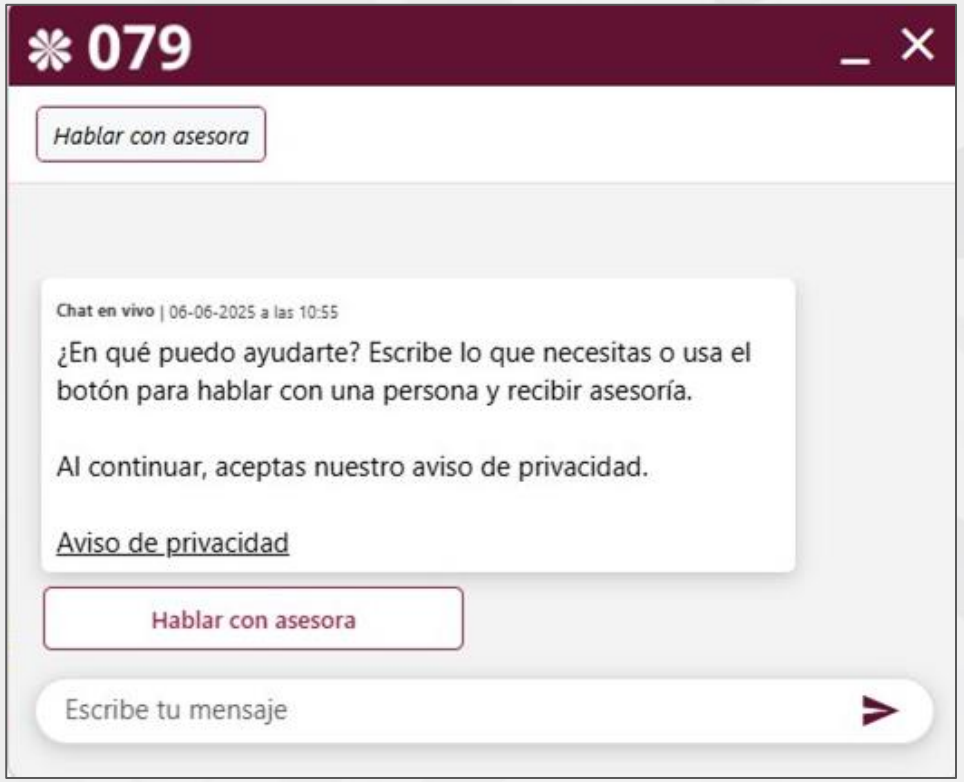
Dr. Carlos Fidel Selva Ochoa
11 de junio de 2025

2025 Año de La Mujer Indígena

Chatbot 079

Índice

- Introducción
- Beneficios e Impacto
- Modelo de lenguaje de gran tamaño (LLM)
- Tecnología abierta para soluciones públicas
- Uso de generación de contenido potenciada por búsqueda (RAG)
- Búsqueda semántica y relevancia (Embeddings)
    - Pipeline de interacción inteligente: extracción, búsqueda y generación
- Desarrollo escalable (Kubernetes)



Chatbot 079

Introducción

El acceso a los trámites y servicios públicos ha presentado históricamente desafíos en términos de **simplicidad, coordinación y rapidez**.

Este proyecto surge como una respuesta tecnológica y social para guiar a cualquier ciudadano y **facilitarle el acceso** a los servicios a los que tiene derecho, independientemente de su familiaridad con el proceso.

El portal integra datos oficiales, un buscador semántico avanzado y un chatbot conversacional para que cualquier persona pueda encontrar fácilmente la información necesaria sobre **qué hacer, cómo hacerlo y dónde hacerlo**.

Beneficios e impacto

Para más de 100 millones de personas

El **Chatbot 079** es una herramienta de IA diseñada para optimizar el acceso a servicios públicos y programas sociales en México. Como parte de la transformación digital, integrará datos oficiales para responder consultas de forma rápida y accesible.

- Plataforma accesible 24/7 desde cualquier dispositivo, sin necesidad de trasladarse a oficinas.
- Integra más de 30,000 trámites con información estructurada, actualizada y clara.
- Reduce la burocracia al centralizar fuentes y automatizar respuestas.
- Promueve la inclusión digital y reduce la brecha en el acceso a servicios.
- Ayuda a combatir la corrupción al hacer transparente el procedimiento de cada trámite.

2025
Año de
La Mujer
Indígena

Modelo del lenguaje de gran tamaño (LLM)

Un motor que entiende el lenguaje cotidiano

El corazón del buscador y el chatbot es un modelo de lenguaje de gran tamaño (LLM), entrenado con lenguaje administrativo y jurídico mexicano.

Gracias a este modelo:

- El sistema entiende preguntas en lenguaje natural como “quiero renovar mi pasaporte”.
- Puede explicar diferencias entre trámites similares o complejos.
- Da respuestas más cercanas a cómo piensa un usuario real.

Actualmente utilizamos una arquitectura basada en dos LLMs especializados:

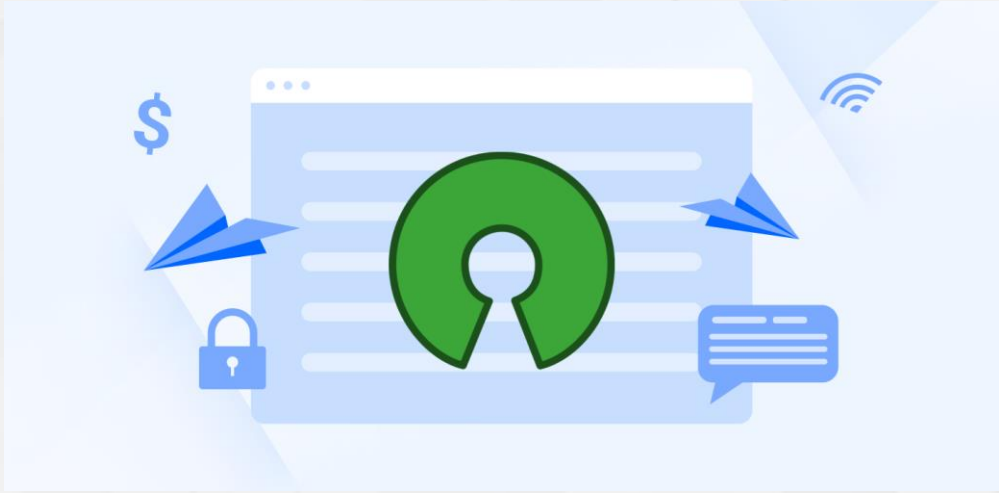
- LLM Extractor: interpreta la necesidad del usuario y la traduce a parámetros semánticos para la búsqueda.
- LLM Formateador: organiza y presenta los resultados de forma clara, útil y adaptada al lenguaje ciudadano.

Todo esto se ejecuta con el modelo Qwen 3-8B, un modelo eficiente, multilingüe y de código abierto, desplegado de forma local a través de Ollama, lo que garantiza soberanía tecnológica, tiempos de respuesta bajos e integración con sistemas gubernamentales.

Tecnología abierta para soluciones públicas

Licencia de uso libre

El desarrollo de esta plataforma se sustenta en software libre (FOSS), lo que permite:



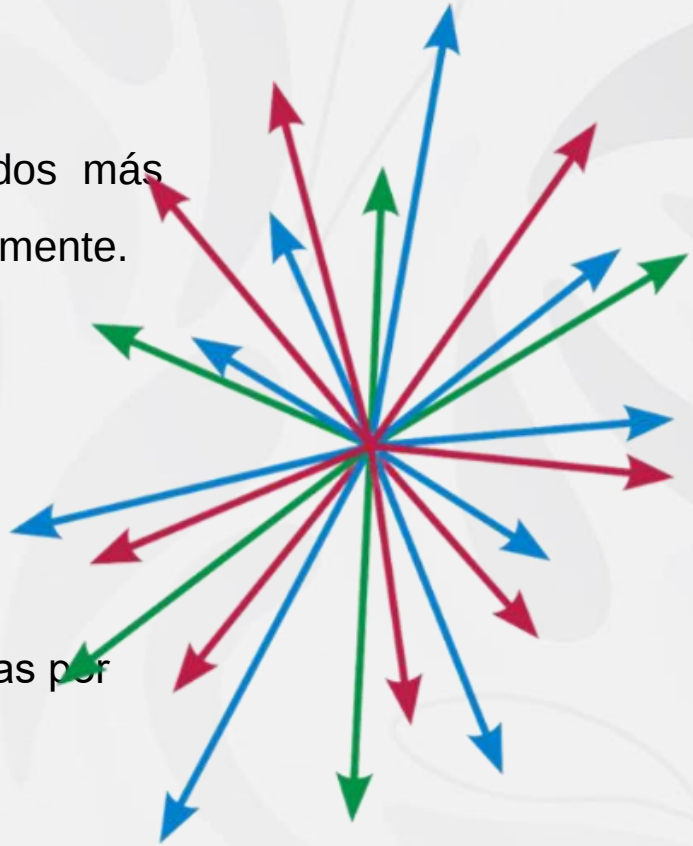
- Transparencia en el código y en las decisiones del sistema.
- Replicabilidad por gobiernos estatales o municipales.
- Ahorro en licencias y dependencias externas.
- Esto permite construir un ecosistema soberano de innovación pública, donde cualquier entidad pueda adaptar y mejorar la solución.

Uso de generación de contenido potenciada por búsqueda

RAG: precisión aumentada con datos verificados

A diferencia de modelos tradicionales que generan texto sin fuentes, este sistema utiliza **RAG (Retrieval-Augmented Generation)**:

Nuestro RAG incluye un paso de re-ranking semántico para priorizar los resultados más alineados con la intención del usuario, incluso si las palabras clave no coinciden exactamente.



- Antes de responder, busca en una base oficial vectorizada.
- Luego genera una respuesta usando sólo información relevante.
- Esto asegura que la IA no “alucine” y que las respuestas estén siempre respaldadas por datos reales.

Preprocesamiento de los datos de contexto a un espacio latente

Representación inteligente de los trámites

Cada trámite se transforma en un vector semántico: un punto en un espacio multidimensional donde se agrupan por similitud. Este espacio latente permite representar tanto los trámites como las preguntas de los usuarios.

Este proceso permite:

- Buscar no solo por palabras exactas, sino por intención del usuario.
- Hacer sugerencias más humanas (por ejemplo: “quiero registar a mi hijo” sugiere trámites de nacimiento).
- Responder aunque el usuario no conozca el nombre exacto del trámite.

Además, este mismo espacio vectorial es el que se utiliza en la etapa de re-ranking semántico, donde el sistema compara de forma precisa la intención del usuario con los resultados obtenidos vía API. Así se priorizan los trámites más relevantes y útiles, mejorando la calidad final de la respuesta.

Desarrollo escalable (Kubernetes)

Infraestructura lista para crecer

El backend del sistema está orquestado con **Kubernetes**, lo que permite:

- Alta disponibilidad para millones de usuarios simultáneos.
- Escalamiento automático según demanda.
- Monitoreo de servicios en tiempo real.
- Además, cada componente (API, vectorizador, frontend, chatbot) puede actualizarse de forma independiente, reduciendo tiempos de mantenimiento y errores.