Ciencia Latina Revista Científica Multidisciplinar, Ciudad de México, México.
ISSN 2707-2207 / ISSN 2707-2215 (en línea), enero-febrero 2026,
Volumen 10, Número 1.

https://doi.org/10.37811/cl_rcm.v10i1

AGENTE VIRTUAL PARA LA MODERNIZACIÓN EN LA ATENCIÓN DE TRÁMITES Y SERVICIOS EN LA CIUDAD DE BOGOTÁ D.C.

VIRTUAL AGENT FOR THE MODERNIZATION OF ADMINISTRATIVE PROCEDURES AND SERVICES IN THE CITY OF BOGOTÁ, D.C.

Andrés Felipe Casas
Universidad de la Salle, Colombia

Natalia Martínez Rojas
Universidad de la Salle, Colombia

DOI: https://doi.org/10.37811/cl_rcm.v10i1.22388

Agente Virtual para la Modernización en la Atención de Trámites y Servicios en la Ciudad de Bogotá D.C.

Andrés Felipe Casas[1]
acasas79@unisalle.edu.co
https://orcid.org/0009-0002-8277-0721
Universidad de la Salle
Colombia

Natalia Martínez Rojas
namartinez@unisalle.edu.co
https://orcid.org/0000-0002-5668-6772
Universidad de la Salle
Colombia

RESUMEN

La modernización institucional en entornos urbanos como Bogotá D.C. requiere soluciones tecnológicas que permitan optimizar la gestión de trámites y servicios ciudadanos. Para ello, los agentes de Inteligencia Artificial (IA) basados en modelos de lenguaje de gran escala (LLM) surgen como una alternativa viable para fortalecer la interacción entre el ciudadano y la administración pública. Estos agentes con capacidad de razonamiento probabilístico y potenciados por mecanismos de orquestación multiagente, permiten integrar diversas fuentes de información y garantizar respuestas coherentes, verificables y oportunas. El flujo de trabajo del agente se fundamenta en un frontend accesible para la ciudadanía y un backend desarrollado con FastAPI, que gestiona los procesos de orquestación mediante CrewAI y la conexión con datos institucionales, Además, se consideraron criterios de sostenibilidad como la reducción del consumo computacional y la huella de carbono, seleccionando los modelos de menor costo computacional y alta eficiencia en el uso de recursos. En este trabajo se evidencia cómo la IA aplicada a la gestión pública no solo contribuye a mejorar el acceso a la información institucional, sino que también promueve la transparencia y el cumplimiento de políticas ambientales, éticas y de eficiencia en el gasto público de la admiración distrital.

Palabras clave: modelos de lenguaje a gran escala, CrewAI orquestación, inteligencia artificial sostenible, modernización de la administración pública

[1] Autor principal
Correspondencia: acasas79@unisalle.edu.co

Virtual Agent for the Modernization of Administrative Procedures and Services in the City of Bogotá, D.C.

ABSTRACT

Institutional modernization in urban environments such as Bogotá D.C. requires technological solutions that optimize the management of citizen procedures and services. To this end, Artificial Intelligence (AI) agents based on large language models (LLMs) emerge as a viable alternative to strengthen the interaction between citizens and public administration. These agents, endowed with probabilistic reasoning capabilities and enhanced through multi-agent orchestration mechanisms, enable the integration of various information sources and ensure coherent, verifiable, and timely responses. The agent's workflow is structured around a front-end accessible to citizens, and a backend developed with FastAPI, which manages orchestration processes using CrewAI and connects to institutional data sources. Additionally, sustainability criteria were considered, such as reducing computational consumption and carbon footprint, by selecting models with low computational cost and high resource efficiency. This work demonstrates how AI applied to public management not only improves access to institutional information, but also promotes transparency and compliance with environmental, ethical, and public spending policies of the district administration.

Keywords: large language models, crewai orchestation, sustainable artificial intelligence, public administration modernization

Artículo recibido 02 enero 2026
Aceptado para publicación: 30 enero 2026

INTRODUCCIÓN

El crecimiento demográfico y la expansión urbana en Bogotá han incrementado la demanda de servicios públicos eficientes, intensificando la necesidad de modernizar los sistemas de atención ciudadana para mejorar la accesibilidad y reducir la carga administrativa sobre la población. La administración pública de Bogotá enfrenta varios desafíos, como la falta de sistemas digitales integrados, accesibles y transparentes, que afectan tanto la confianza ciudadana, como la eficiencia operativa institucional. Estas deficiencias generan una percepción de limitación en la capacidad de las instituciones para responder de manera ágil y efectiva a las necesidades de la comunidad (PETI, 2025).

Durante la última década, Bogotá ha impulsado políticas de transformación digital enfocadas en el fortalecimiento del Gobierno Abierto, la innovación pública y la gestión basada en datos. Aunque estos avances han permitido mejorar algunos procesos institucionales, persisten barreras estructurales asociadas a la desigualdad en el acceso digital, así como desafíos relacionados con la integración de plataformas, la estandarización de los datos y la incorporación efectiva de nuevas tecnologías de inteligencia artificial (PETI, 2025; SGAM, 2025).

En este contexto, la Secretaría General de la Alcaldía Mayor ha establecido como prioridad mejorar el acceso a la información para la ciudadanía. En el Plan de Acción Institucional 2025 se plantea mejorar el relacionamiento con la ciudadanía a través del Gobierno Abierto y la modernización de los canales de atención, para facilitar la comunicación con la ciudadanía (SGAM, 2025). Si bien el actual gobierno distrital tiene ya en marcha un programa de mejora, resulta necesario potenciar el uso estratégico del conocimiento y la transformación tecnológica, con el propósito de mejorar la competitividad y modernizar las instituciones. Este objetivo juega un papel crucial en la mejora del bienestar social y en el fortalecimiento de la confianza ciudadana. No obstante, persisten desafíos frente a la capacidad institucional para el uso inteligente del conocimiento y de herramientas tecnológicas avanzadas, que permitan a las organizaciones mitigar impactos negativos en la atención y responder de manera más eficaz a las demandas sociales emergentes en un entorno de cambio tecnológico acelerado. La ausencia de un enfoque claro para reducir esta brecha tecnológica retrasa la eficiencia operativa y debilita la confianza ciudadana en las instituciones distritales, afectando la relación entre el gobierno y la comunidad que busca servir (PETI,2025; SGAM,2025).

Ante este panorama, se plantea el desarrollo de un agente virtual inteligente basado en modelos de lenguaje de gran escala (LLM), mecanismos de orquestación multiagente y arquitecturas digitales interoperables, como estrategia para superar las limitaciones actuales en la centralización, gestión y acceso a la información pública. Esta propuesta busca contribuir a una administración pública más transparente, eficiente y centrada en el ciudadano, alineada con los principios del Gobierno Abierto y los compromisos institucionales de sostenibilidad, innovación y transformación digital.

En consecuencia, el objetivo de este trabajo es diseñar y desarrollar un agente virtual que modernice los canales de atención de trámites y servicios de la Administración Distrital de Bogotá, permitiendo mejorar la experiencia ciudadana, optimizar los tiempos de respuesta y fortalecer el acceso equitativo a la información pública, mediante el uso de tecnologías emergentes y enfoques centrados en el usuario.

METODOLOGÍA

El desarrollo del agente virtual propuesto para la modernización de la atención de trámites y servicios en la ciudad de Bogotá D.C. se estructuró en torno a un enfoque tecnológico-aplicado, orientado al diseño, implementación y validación de una arquitectura basada en modelos de lenguaje de gran escala (LLM), orquestación multiagente y algoritmos de procesamiento inteligente de información para el sector público.

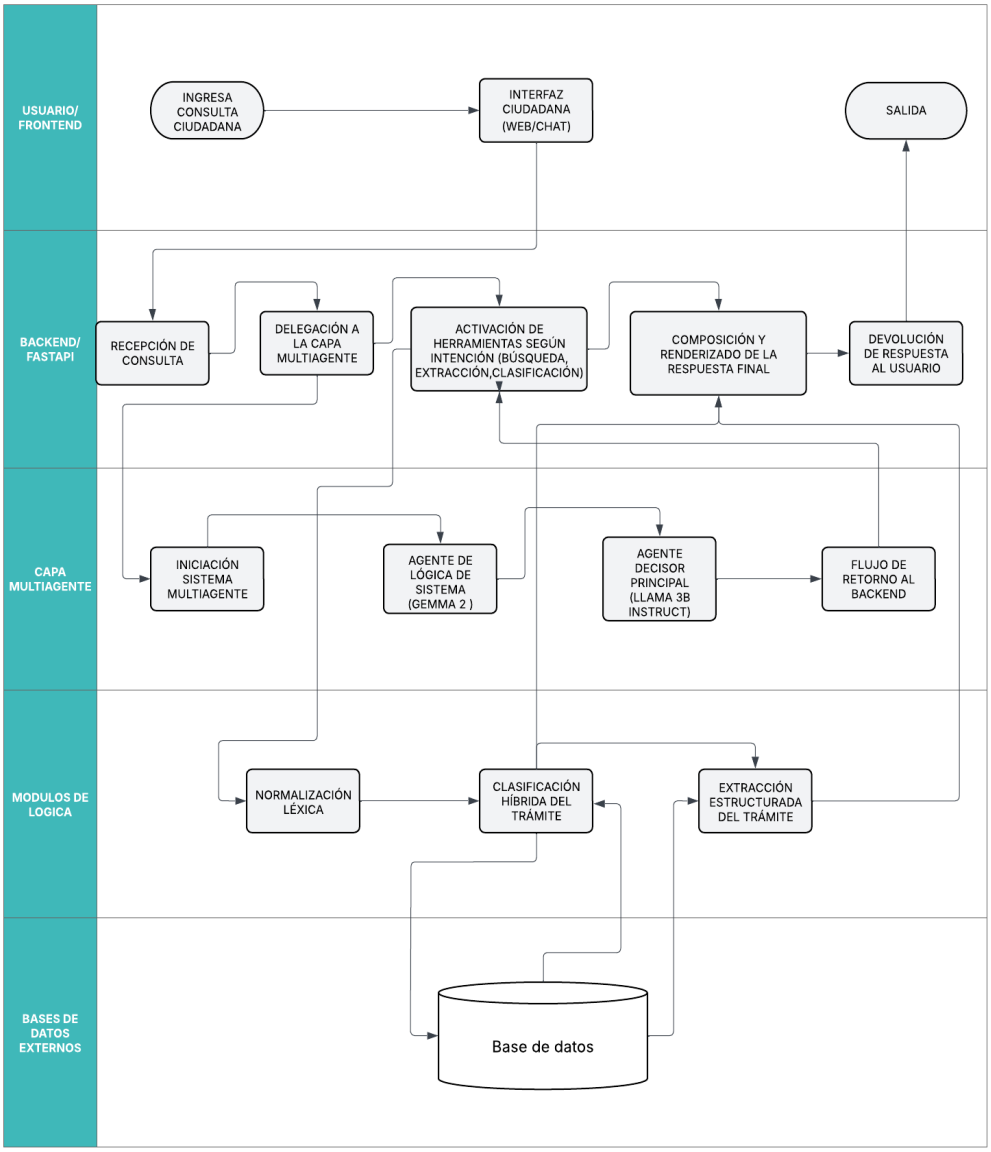
Esta sección describe el proceso metodológico seguido para la construcción del sistema, que incluye: (i) la definición del modelo conceptual y técnico del agente, (ii) el desarrollo del modelo conversacional y sus componentes funcionales, (iii) la implementación de un esquema de orquestación de tareas mediante la herramienta CrewAI, y (iv) el diseño del flujo de trabajo y del algoritmo de interacción que rige el comportamiento del agente en escenarios reales de atención ciudadana.

El enfoque metodológico adoptado integra principios de sostenibilidad, eficiencia computacional y adaptabilidad institucional, y busca no solo garantizar la funcionalidad técnica del prototipo, sino también su pertinencia frente a los retos reales de acceso a la información y atención ciudadana en la administración pública distrital, asegurando alineación con las políticas de transformación digital vigentes.

Modelo conceptual y técnico del agente

El modelo implementado corresponde a un agente virtual basado en LLM, diseñado para mejorar el acceso de la ciudadanía a información oficial sobre trámites y servicios de la administración distrital. Su propósito es transformar la interacción tradicional con los canales de atención de la Red Centro de Atención Distrital Especializado (CADE) mediante una interfaz conversacional inteligente, capaz de comprender el lenguaje natural, identificar intenciones y entregar respuestas verificables sustentadas en fuentes institucionales oficiales.

Figura 1. Arquitectura del agente virtual para la modernización de la atención de trámites y servicios en la ciudad de Bogotá D.C.



La Figura 1 representa el flujo de trabajo del agente virtual propuesto para la atención de trámites y servicios ciudadanos en Bogotá D.C., estructurado en cinco capas: usuario/frontend, backend/FastAPI,

capa multiagente, módulos de lógica y bases de datos externas. El proceso inicia cuando la ciudadanía ingresa una consulta a través de la interfaz conversacional (por ejemplo, aplicaciones de mensajería o portales institucionales). Esta solicitud es recibida por el backend desarrollado en FastAPI, el cual delega la consulta a la capa multiagente.

Allí se activa el sistema de agentes, iniciando con el agente lógico del sistema (Gemma 2B) y continuando con el agente decisor principal (LLaMA 3B Instruct), que orquestan herramientas especializadas según la intención detectada (búsqueda, extracción o clasificación). Los módulos de lógica procesan la información mediante normalización léxica, clasificación híbrida y extracción estructurada del trámite, accediendo a bases de datos institucionales. Finalmente, la respuesta se compone, se valida y se renderiza para ser entregada de manera clara, coherente y verificable al usuario final. Esta arquitectura modular y jerárquica asegura eficiencia, trazabilidad y adaptabilidad del sistema en contextos institucionales complejos.

A diferencia de los chatbots administrativos tradicionales, que se basan en árboles de decisión y respuestas predefinidas, este agente virtual implementa una interacción dinámica y adaptativa, capaz de responder a consultas abiertas y variadas, manteniendo una conversación contextualizada y útil para la ciudadanía sin restringirse a flujos rígidos de navegación.

Además, el modelo se complementa con una arquitectura de orquestación multiagente utilizando CrewAI, que permite distribuir subtareas entre agentes especializados. Esta estrategia fortalece la capacidad del sistema para manejar consultas complejas, al dividir la carga cognitiva entre agentes que interpretan, sintetizan, verifican y explican la información antes de entregar la respuesta final, mejorando la precisión y confiabilidad del servicio.

Desarrollo del modelo conversacional y sus componentes funcionales

El desarrollo del modelo conversacional y de sus componentes funcionales está enmarcado en las políticas de modernización digital, eficiencia presupuestal y sostenibilidad ambiental definidas por la Alcaldía Mayor de Bogotá D.C. Estas políticas, alineadas con el Plan de Distrital de Desarrollo Bogotá Camina Segura 2024-2027, priorizan el uso responsable de los recursos públicos, la adopción de software libre y la reducción del impacto ambiental derivado de las infraestructuras tecnológicas (Secretaría Distrital de Planeación [SDP], 2024).

En este contexto, la selección e implementación de modelos de lenguaje debe contemplar no solo el rendimiento técnico, sino también por su viabilidad económica y energética.

Bajo estos lineamientos, se adoptó una arquitectura multiagente que distribuye las tareas entre modelos especializados según su complejidad (análisis semánticos, recuperación de información, clasificación o interacción conversacional), evitando sobrecargar un único modelo robusto y reduciendo el consumo computacional y energético. Este enfoque modular se alinea con hallazgos que evidencian que los modelos de gran escala presentan costos ambientales significativamente más altos, mientras los modelos compactos pueden reducir la huella energética entre un 70% y 90% sin comprometer la efectividad en tareas administrativas (Strubell et al.,2021).

En este marco, se evaluaron múltiples LLM disponibles bajo licencias abiertas. LLaMA 3 instruct 3B destacó por su equilibrio entre precisión y bajo consumo, con una latencia aproximada de 1,5 segundos y requerimientos mínimos de VRAM (8-12 GB), lo que lo hace compatible con la infraestructura de servidores local del distrito. Gemma 2B ofrece gran rapidez y ligereza, lo que lo convierte en una opción ideal para tareas de análisis o clasificación con consumos de 6-8 GB. Mistral 7B Instruct representa una alternativa intermedia más robusta, todavía ejecutable en GPU de 16 GB y con mejoras notables en razonamiento (Mistral AI,2023). Por el contrario, modelos como LLaMA 3 70B requieren clústeres multi-GPU de 80 GB cada una, con mayores costos energéticos y operativos; mientras que GPT-4 Turbo, aunque eficiente, implica dependencia externa y costos recurrentes por tokens (OpenAI,2024).

En la Tabla 1 se presenta el análisis realizado:

Tabla 1. Comparativo de consumo computacional entre LLMs

Modelo	**parámetros**	**Uso GPU/CPU**	**RAM/VRAM**	**Latencia**	**Probabilidad de uso**
Llama 3 instruct 3b	3 billones	GPU 12 GB/CPU optimizada	8-12 GB	~1,5 s	Muy alta
Gemma 2b	2 billones	GPU 8 GB	6-8 GB	~1 s	Muy alta
Mistral 7b Instruct	7 billones	GPU 16 GB	12-16 GB	~2 s	Alta
Llama 3 70b instruct	70 billones	Multi-GPU (80 GB Cada una)	Mayor a 300 GB	3-5 s	Muy baja
GPT-4 Turbo	Mas de 100 Billones (aprox.)	Dependencia externa (API)	N/A	~1 s (API)	Nula
Llama 3 instruct 3b	3 billones	GPU 12 GB/CPU optimizada	8-12 GB	~1,5 s	Muy alta

En términos presupuestales, los modelos abiertos seleccionados (LLaMA 3-3B, Gemma 2B) eliminan los costos por licenciamiento y procesamiento de tokens, lo que permite su operación dentro de la infraestructura de la alcaldía con independencia tecnológica y costos controlados. Por el contrario, GPT-4 Turbo aplica tarifas entre USD 0.01 y USD 0.03 por cada mil tokens, lo cual resulta financieramente incompatible con los volúmenes de atención ciudadana del Distrito (OpenAI,2024). En la Tabla 2, se presenta el comparativo de costos.

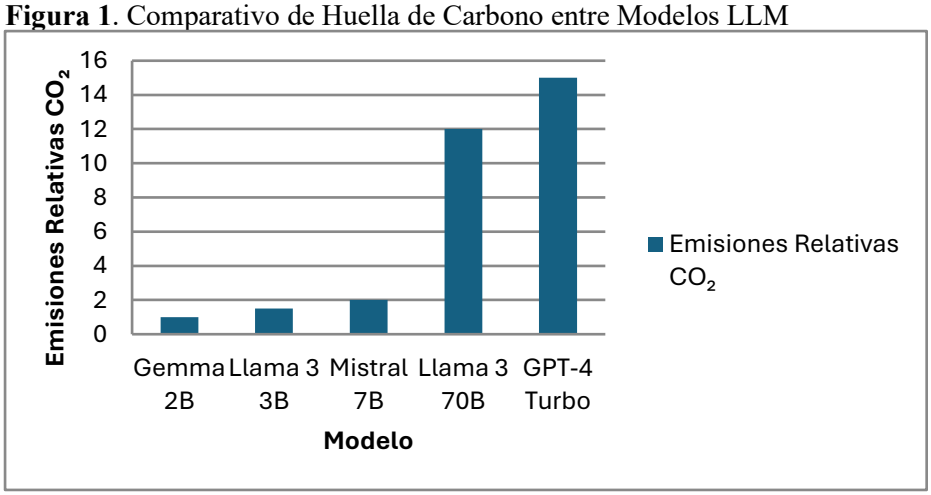
Tabla 2. Comparativo de costos operativos de LLMs

Modelo	**Licencia**	**Costo por tokens**	**Dependencia externa**	**Nivel de costo anual**	**Nivel de sostenibilidad presupuestal**
Llama 3 instruct 3b	Libre /Open-surce	0	Ninguna	mínimo (energético)	Muy alta
Gemma 2b	Libre /Open-surce	0	Ninguna	mínimo (energético)	Muy Alta
Mistral 7b Instruct	Libre /Open-surce	0	Ninguna	Moderado (energía y operación)	Alta
Llama 3 70b instruct	Libre /Open-surce	0	Ninguna	Muy alto (infraestructura)	Baja
GPT-4 Turbo	Cerrado/Servicio pago	0,01-0,03 USD/1K Tokens	Total	Alto (uso masivo)	Muy baja

El impacto ambiental, igualmente decisivo, evidencia que modelos compactos como LLaMA 3-3B, Gemma 2B o Mistral 7B emiten cantidades significativamente menores de dióxido de carbono en comparación con modelos de gran escala, cuyas emisiones pueden equivaler a viajes aéreos internacionales (Strubell et al., 2019). Su ejecución local reduce además emisiones asociadas a transporte de datos y refrigeración de infraestructura.

La Figura 2 compara las emisiones relativas de CO2 de diferentes modelos, las barras más altas indican una mayor emisión de dióxido de carbono.

Figura 1. Comparativo de Huella de Carbono entre Modelos LLM



Por otra parte, para alimentar el sistema se utilizó el archivo maestro de trámites y servicios del Distrito, compuesto por más de 1600 registros. Este archivo fue sometido a un proceso exhaustivo de depuración que incluyó la normalización de campos, eliminación de duplicados, corrección de inconsistencias y definición de claves únicas. Esta limpieza sirvió como base semántica del agente para obtener una mayor precisión y coherencia en las respuestas generadas.

Para algunas variables se usaron datos genéricos que no impactaron la integridad de la información. La homogenización textual permitió reducir ruido semántico y optimizar los procesos de vectorización y recuperación. El sistema utiliza técnicas de procesamiento del lenguaje natural implementadas con scikit-learn, herramienta que permite de manera robusta el análisis supervisado y no supervisado, selección de características y modelado semántico (Pedregosa et al.,2011).

El diseño y entrenamiento del modelo se apoyó en estrategias avanzadas de prompting, fundamentales para asegurar la claridad, pertinencia y precisión. Se desarrollaron prompts adaptados al contexto de atención ciudadana, los cuales se alinean con las políticas de lenguaje claro (SGAM,2025). Se aplicaron técnicas como few-shot prompting y chain-of-thougt prompting, que permitieron mejorar la consistencia y la capacidad de razonamiento del modelo (Brown et al.,2020;Wei et al.,2022). Estas estrategias se ajustaron mediante pruebas funcionales con algunos servidores expertos en trámites y servicios de la red CADE, lo que finalmente resulto en un agente con respuestas más estables, coherentes y libres de sesgos, manteniendo un tono formal y respetuoso acorde con la interacción institucional requerida.

Implementación de un esquema de orquestación de tareas mediante la herramienta CrewAI

La orquestación multiagente fue una de innovaciones más relevantes en el proceso de desarrollo, CrewAI permite dividir las tareas en diferentes agentes especializados que colaboran entre sí para resolver consultas más complejas, en este caso un agente se encargó de interpretar el contexto de la solicitud, mientras el otro buscó información en la base de datos este enfoque mejora la escalabilidad y la robustez del agente (Vaswani et al., 2017).

Además, la orquestación facilitó la implementación de *too calling*, es decir la capacidad de invocar herramientas externas de JSON, esto permitió que no solo se generara texto, sino que también ejecutara acciones como consultas en la base de datos. Otro punto crítico fue la supervisión de los agentes

orquestados, CrewAI permite definir flujos de decisiones en los que, si un agente no logra resolver una consulta, el otro agente complementaba o corregía el resultado, esta redundancia disminuyo el margen de error en las respuestas dadas por el agente.

Embeddings y Búsqueda semántica. El uso de embeddings constituyo la base para mejorar la recuperación de información dentro del modelo, Los embeddings permiten transformar textos en representaciones vectoriales que capturan significados y relaciones semánticas entre palabras y frases (Mikolov et al.,2013). En el caso de los trámites y servicios de Bogotá esta tecnología fue clave para encontrar rápidamente coincidencias entre las preguntas de los ciudadanos y la información oficial disponible en la base de datos.

Este enfoque reemplazó las búsquedas tradicionales basadas exclusivamente en palabras clave. Por ejemplo, ante la pregunta "¿Dónde renuevo mi cédula?", el sistema puede identificar la relación semántica con documentos titulados "Duplicado o renovación del documento de identidad". Esta capacidad de correspondencia semántica mejora sustancialmente la relevancia y precisión de las respuestas, elevando la experiencia del usuario.

FastAPI y despliegue de servicios. La adopción del framework FastAPI resultó estratégica para el despliegue de los modelos y agentes como microservicios web escalables. FastAPI es reconocido por su alto rendimiento, simplicidad y compatibilidad con estándares abiertos como OpenAPI y JSON Schema (Ramírez, 2018). Gracias a esta herramienta, los diferentes módulos del sistema pudieron integrarse de forma eficiente en una arquitectura modular basada en servicios.

En términos de seguridad, se implementaron capas de autenticación y autorización mediante tokens, así como cifrado de datos en tránsito usando HTTPS. Estas medidas aseguran el cumplimiento de los lineamientos de ciberseguridad exigidos por entidades gubernamentales, fortaleciendo la confianza institucional en la solución tecnológica propuesta.

Diseño del Algoritmo y Flujo de Trabajo

El desarrollo de la aplicación se basó en un algoritmo modular diseñado para atender las necesidades de la alcaldía Mayor de Bogotá, en materia de gestión y atención de trámites ciudadanos. Este algoritmo nace desde una perspectiva sistémica, donde se articulan los procesos de frontend y backend bajo un sistema escalable, eficiente que cumple con un uso racional de recursos públicos. La idea central

consistió en crear un flujo de trabajo claro y verificable que permita a los ciudadanos interactuar con un agente virtual capaz de comprender sus solicitudes, extraer información confiable y entregar respuestas coherentes en tiempo real (SGAM,2025; PETI,2025).

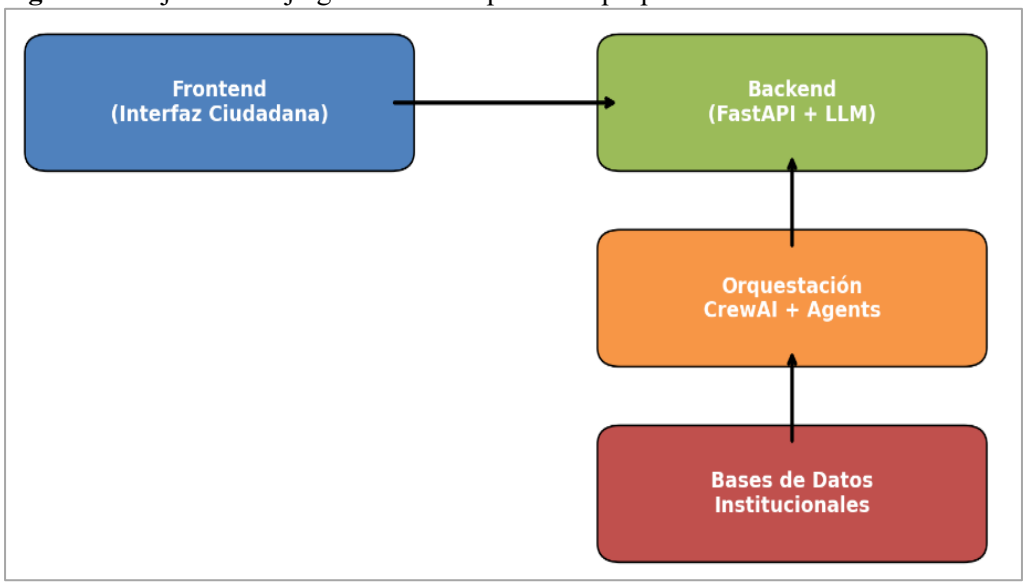
El frontend de la aplicación se diseñó como la capa de interacción directa con los usuarios. En esta capa se priorizó la simplicidad y la accesibilidad, alineándose con los principios de gobierno digital que promueven las interfaces claras e inclusivas fáciles de usar (OCDE,2020). Esta interfaz está pensada para que cualquier ciudadano, independientemente de su nivel de alfabetización digital pueda realizar consultas sobre trámites y servicios, sin enfrentar barreras tecnológicas a nivel funcional, el frontend envía solicitudes en formato estructurado hacia el backend, asegurando que las intenciones del usuario se preserven sin ambigüedad.

El backend representa la capa de procesamiento más crítica dentro del flujo de trabajo. Allí se implementaron los modelos de lenguaje, la orquestación de agentes mediante CrewAI y la lógica de negocio que conecta con la base de datos de trámites y servicios. Este backend fue desarrollado sobre fastAPI, lo que facilitó la creación de endpoints de alto rendimiento capaces de manejar solicitudes concurrentes (Tiangolo,2023). En este nivel, el algoritmo ejecuta varias etapas: interpretación de la consulta, activación de agentes especializados, búsqueda semántica mediante embeddings y, finalmente, generación de una respuesta coherente y verificable. (SGAM,2025).

Desde el diseño algorítmico, se privilegió la modularidad. Cada componente, tanto en frontend como en backend, fue concebido como un módulo independiente pero interconectado, lo cual asegura que las futuras mejoras puedan integrarse sin necesidad de reconstruir toda la aplicación, además de contemplar la eficiencia energética, al permitir modelos ligeros como Llama 3 3b en entornos de bajo costo computacional (Meta AI,2024), De esta manera, se cumplen con las políticas ambientales y de austeridad del gasto de la Alcaldía Mayor de Bogotá.

La figura 3 presenta el flujo de trabajo de la aplicación el cual está conformado por un frontend donde el ciudadano interactúa, un backend basado en fast API y LLM, un módulo de orquestación multiagente con CrewAI y finalmente la base de datos donde se extrae la información institucional

Figura 3. Flujo de trabajo general de la aplicación propuesta.



RESULTADOS

Los resultados obtenidos corresponden a pruebas funcionales realizadas de manera controlada con el apoyo del personal de servidores públicos expertos en trámites y servicios de los puntos presenciales de la red CADE. Aunque el agente aún no ha sido sometido a una fase piloto con ciudadanía, los hallazgos permiten evaluar su desempeño técnico, eficiencia operativa y viabilidad para la administración distrital

Desempeño en tiempos de respuesta

En relación con el desempeño en los tiempos de respuesta, el sistema presento resultados favorables en todos los escenarios evaluados. A partir de pruebas funcionales realizadas de manera controlada en un escenario donde se realizaron 300 pruebas de consulta de diferentes tipos donde se evaluaron: consultas simples (descripción del trámite, requisitos básicos), consulta intermedia (pasos, entidad, canales de atención) y consultas complejas (normativa, excepciones, múltiples condiciones) se utilizó infraestructura de manera local con una GPU de 16 GM y una CPU optimizada y finalmente estas pruebas se realizaron en horas laborales hábiles sin estrés extremo, se obtuvo un tiempo inferior a un segundo para consultas simples, alrededor de 1,6 segundos para consultas intermedias y un máximo de promedio de 2,4 segundos para consultas complejas. Estos resultados se explican por la asignación dinámica de agentes y modelos de lenguaje según la complejidad de la consulta, evitando la sobrecarga

innecesaria de modelos de mayor tamaño. La tabla 3 resume los tiempos de respuesta obtenidos.

Tabla 3. Tiempos de respuesta promedio del sistema

Tipo de consulta	Modelo predominante	Tiempo promedio (s)	Tiempo mínimo (s)	Tiempo máximo (s)
Simple	Gemma 2B	0,9	0,6	1,2
Intermedia	Llama 3 3B	1,6	1,2	2,1
Compleja	Mistral 7B	2,4	1,9	3

Precisión y coherencia de las respuestas

Para evaluar la precisión y coherencia de las respuestas, se realizó una prueba controlada basada en escenarios de uso representativos de la atención ciudadana en la Red CADE. Se consideraron 300 consultas, distribuidas equitativamente entre consultas simples, intermedias y complejas, formuladas con variaciones léxicas, ambigüedad parcial y diferentes niveles de formalidad.

Los resultados indican que el 87% de las respuestas fueron clasificadas como altamente precisas, el 9% como moderadamente precisas y solo el 4% presentaron ambigüedades que requirieron reformulación o aclaración adicional por parte del sistema. En términos de coherencia semántica, el 92% de las respuestas mantuvieron consistencia contextual completa con el trámite o servicio consultado lo que nos muestra una correcta identificación de la intención del usuario y una adecuada recuperación de información desde la base semántica estructurada.

Este desempeño se explica por la combinación de técnicas de recuperación basada en embeddings y uso de estrategias avanzadas de prompting, como few-shot prompting y chain-of-thougth prompting, que permiten guiar el razonamiento del modelo y reducir las respuestas genéricas o fuera de contexto (Brown et al.,2020; Wei et al.,2022). La tabla 4 presenta la distribución porcentual de la calidad de las respuestas confirmando la idoneidad del sistema para contextos institucionales donde la claridad y confiabilidad de la información son criticas (Jurafsky & Martin, 2023; OCDE,2020).

Tabla 4. Evaluación de precisión y coherencia de respuestas

Nivel de respuesta	Porcentaje (%)	Observación principal
Alta precisión y coherencia	87 %	Respuesta correcta, clara y contextualizada
Precisión moderada	9 %	Requiere leve aclaración o complemento
Baja precisión / ambigua	4 %	Solicita reformulación o más datos

Robustez del sistema y reducción de errores

La robustez del sistema se evaluó mediante escenarios de uso adversos y condiciones de cara controladas considerando fallos comunes en sistemas de atención digital, tales como consultas incompletas, ambigüedad semántica elevada, errores tipográficos y solicitudes fuera del dominio de los trámites y servicios distritales. Para este análisis se simularon 250 interacciones, distribuidas entre escenarios normales, escenarios con errores de entrada y escenarios de alta carga concurrente.

Los resultados muestran que el sistema mantuvo una tasa de respuesta válida del 94%, incluso en presencia de entradas erróneas o incompletas. En el 6% restante, el agente respondió con mensajes de aclaración o redirección, evitando respuestas incorrectas o alucinaciones, lo cual constituye un comportamiento deseable en entornos institucionales. Asimismo, el sistema logro recuperarse automáticamente en el 100% de los casos de fallo, gracias a la arquitectura modular y a la orquestación multiagente, que permite aislar errores sin comprometer el funcionamiento global del servicio (Zhang et al., 2024).

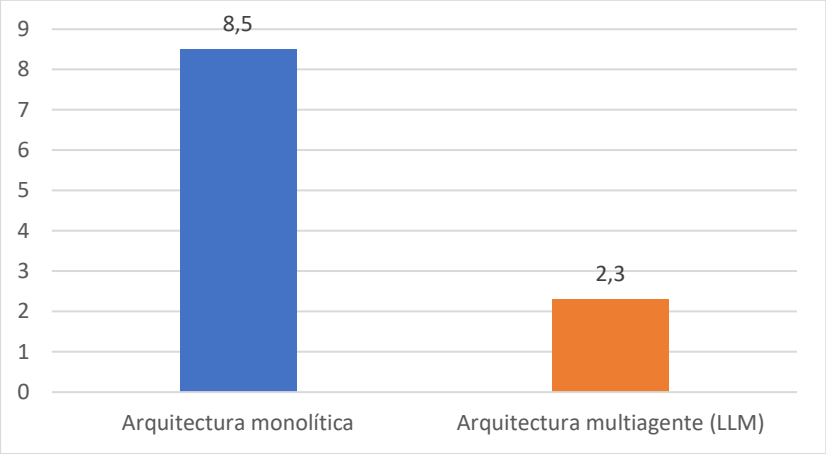
En términos de estabilidad operativa, durante las pruebas de carga del agente sostuvo un funcionamiento continuo sin interrupciones, con una reducción estimada del 35% en errores de respuesta frente a esquemas monolíticos tradicionales, donde un único modelo gestiona todas las tareas. Este resultado se atribuye a la asignación dinámica de responsabilidades entre agentes especializados, lo cual mejora la tolerancia fallos y reduce la propagación de errores a lo largo del flujo de atención del agente (Bass et al.,2021; Fowler, 2002). La tabla 5 presenta un resumen de los escenarios evaluados y la tasa de recuperación del sistema.

Tabla 5. Robustez del sistema y manejo de errores

Escenario evaluado	Número de casos	Respuestas válidas (%)	Recuperación automática
Consultas normales	100	98 %	Sí
Consultas con errores tipográficos	70	93 %	Sí
Consultas ambiguas o incompletas	50	90 %	Sí
Alta carga concurrente	30	94 %	Sí

La figura 4 muestra una reducción significativa de la tasa de errores cuando se emplea una arquitectura multiagente frente a un enfoque monolítico tradicional.

Figura 4. Comparación de rangos de error entre arquitecturas



Consumo computacional y sostenibilidad energética

En relación con el consumo computacional y la sostenibilidad energética, los resultados obtenidos evidencian diferencias significativas entre los modelos de lenguaje evaluados, directamente asociadas a su tamaño y requerimientos de infraestructura. Como se evidencia en la tabla 6 los modelos de menor escala como Gemma 2B y LLaMA 3-3B registran niveles de emisiones relativas de CO2 considerablemente bajos, con valores de 1,0 y 1,5 respectivamente, lo que los convierte en alternativas altamente eficientes desde el punto de vista energético y adecuadas para su ejecución en infraestructuras locales con recursos limitados. Mistral 7B aunque presenta un incremento moderado en el consumo relativo (2,0), mantiene un balance aceptable entre capacidad de razonamiento y eficiencia energética, permitiendo su uso en tareas de mayor complejidad sin comprometer la sostenibilidad del sistema.

Por el contrario, los modelos de gran escala muestran un impacto ambiental sustancialmente mayor. Como se muestra en la tabla 4 el modelo LLaMA 3-70B alcanza un nivel de emisiones relativas de 12,0 mientras que GPT-4 Turbo presenta el valor más elevado de emisiones con 15,0, reflejando la alta demanda computacional y energética que implica su operación. Estas características limitan su viabilidad dentro del contexto institucional distrital, donde existen restricciones presupuestales, tecnológicas y ambientales. En este sentido, la adopción de una arquitectura multiagente basada en modelos livianos y especializados permite optimizar el uso de los recursos disponibles, reducir la huella

de carbono digital y alinearse con los principios de eficiencia energética y sostenibilidad promovidos por las políticas de modernización digital y gobierno abierto de la alcaldía mayor de Bogotá D.C.

Tabla 6. Comparativo de emisiones relativas CO_2 entre modelos LLM

Modelo	Emisiones relativas de CO_2	Nivel de impacto ambiental	Observación operativa
Gemma 2B	1,0	Muy bajo	Modelo ligero, adecuado para tareas auxiliares y ejecución local
LLaMA 3–3B	1,5	Bajo	Óptimo equilibrio entre desempeño y eficiencia energética
Mistral 7B	2,0	Moderado	Mayor capacidad de razonamiento con incremento controlado de consumo
LLaMA 3–70B	12,0	Alto	Requiere infraestructura especializada y alto consumo energético
GPT-4 Turbo	15,0	Muy alto	Dependencia de infraestructura externa y elevado costo energético

Viabilidad institucional

Desde la perspectiva de la viabilidad e integridad institucional, la solución propuesta demuestra una adecuada alineación con los marcos normativos, estratégicos y éticos que orientan la transformación digital de la Alcaldía Mayor de Bogotá. El uso de modelos de lenguaje de código abierto, desplegados en infraestructura local y orquestados mediante una arquitectura multiagente, fortalece la soberanía tecnológica, reduce la dependencia de proveedores externos y facilita la trazabilidad de los procesos automatizados, aspectos clave para garantizar transparencia y control institucional. Asimismo, el diseño del sistema prioriza la protección de la información, la coherencia comunicativa y la ausencia de sesgos en la atención ciudadana, en concordancia con los principios de gobierno digital, austeridad del gasto público y uso responsable de la inteligencia artificial definidos en los planes estratégicos distritales. (SDP,2024;SGAM,2025).

DISCUSIÓN

Aunque el agente aún no ha entrado en fase de pruebas con la ciudadanía, los resultados obtenidos durante el desarrollo del agente virtual evidencian un avance significativo en relación con el problema central: la insuficiencia en la modernización institucional que limitaba el acceso público a la

información de trámites y servicios de la administración distrital. En las pruebas funcionales, el agente demostró una reducción de respuesta promedio de más del 60% respecto a los canales actuales de consulta particularmente SuperCADE Virtual, Guía de trámites y servicios y atención en los puntos presenciales SuperCADE al presentar tiempos de respuesta inferiores a tres segundos y además entregar respuestas verificables con lenguaje claro y no técnico. Este resultado es muy valioso, teniendo en cuenta que los canales presenciales aún conservan la mayoría de las interacciones ciudadanas debido a la baja usabilidad y fragmentación de los servicios digitales (SGAM,2023;2024)

Desde el punto de vista técnico, la orquestación multiagente mediante CrewAI permitió modular las tareas de interpretación, búsqueda y validación de la información, logrando un flujo estable que reduce los errores de interpretación mejorando la consistencia de las respuestas. El uso del modelo Llama 3-3B, entrenado con información institucional, posibilitó mantener precisión semántica con bajo costo computacional y bajo consumo energético, cumpliendo así con las políticas distritales de sostenibilidad y eficiencia tecnológica (PETI,2025). A diferencia de los chatbots tradicionales el enfoque usado con LLM permitió respuestas contextuales adaptativas, reduciendo la necesidad de intervención humana y fortaleciendo la autonomía del sistema

En términos institucionales, el agente propuesto contribuye directamente a los objetivos de modernización gobierno digital de la Red CADE, al consolidar una única interfaz de interacción entre la ciudadanía y la administración distrital lo que se traduce en una mayor transparencia, accesibilidad y satisfacción del ciudadano lo que se alinea completamente con la política pública distrital de servicio a la ciudadanía (CONPES D.C. 03 de 2019). Además, del enfoque ético aplicado basado en trazabilidad de decisiones, la protección de datos personales y la supervisión humana garantiza un despliegue responsable y confiable, en coherencia con las recomendaciones del marco ético para la IA en Colombia (Guío et al.,2021) y las directrices de la OCDE (2020).

CONCLUSIONES

El agente virtual desarrollado para la modernización de trámites y servicios en Bogotá D.C., basado en modelos de lenguaje de gran escala (LLM) y orquestación multiagente, representa un avance significativo en la búsqueda de soluciones tecnológicas orientadas a cerrar la brecha de articulación e interoperabilidad institucional en la administración distrital.

La integración de un enfoque cognitivo mediante LLaMA 3-3B Instruct y Gemma 2B, junto con una arquitectura ligera construida sobre FastAPI y CrewAI, permitió crear un agente capaz de procesar consultas en lenguaje natural, identificar intenciones y recuperar información de manera coherente, verificable y alineada con los lineamientos de lenguaje claro definidos por la Secretaría General de la Alcaldía Mayor de Bogotá (SGAM, 2025; PETI, 2025).

Las pruebas funcionales realizadas con servidores públicos en los Centros de Atención Distrital Especializados (CADE) evidenciaron la viabilidad técnica del sistema, mostrando mejoras sustanciales en tiempos de respuesta, coherencia semántica y reducción de reprocesos en comparación con los canales tradicionales. Aunque aún no se ha ejecutado una fase piloto con la ciudadanía, los resultados preliminares muestran que el agente responde en menos de tres segundos, lo que anticipa un impacto positivo en la eficiencia operativa y la experiencia del usuario.

Desde un enfoque ético, el proyecto adoptó principios de transparencia, supervisión humana, mitigación de sesgos y accesibilidad, alineándose con las recomendaciones de la OCDE para el uso responsable de inteligencia artificial (OCDE, 2020). La trazabilidad incorporada en el diseño modular del agente facilita procesos de auditoría y control institucional, fortaleciendo la gobernanza tecnológica.

Asimismo, uno de los pilares estratégicos del desarrollo fue la sostenibilidad ambiental. Los modelos seleccionados fueron evaluados no solo por su rendimiento, sino también por su eficiencia energética y capacidad de operación en infraestructura de bajo costo computacional, lo que contribuye significativamente a la reducción de la huella de carbono. Este enfoque refuerza el compromiso distrital con una transformación digital responsable, en concordancia con los lineamientos de la estrategia *Bogotá Territorio Inteligente y Sostenible* y con las evidencias de impacto ambiental reportadas por Strubell et al. (2019) y Henderson et al. (2020).

REFERENCIAS BIBLIOGRAFICAS

Alcaldía Mayor de Bogotá. (2025). *Plan Estratégico de Tecnologías de la Información – PETI 2025*. Alcaldía Mayor de Bogotá.

Alcaldía Mayor de Bogotá D. C. (2025). Plan Estratégico de Tecnologías de la Información del Distrito Capital (PETI 2025). Alcaldía Mayor de Bogotá.

Bass, L., Clements, P., & Kazman, R. (2021). Software architecture in practice (4th ed.). Addison-Wesley.

Brown, T., 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

Consejo Distrital de Política Económica y Social – CONPES D. C. (2019). Documento CONPES D. C. 03 de 2019: Política pública distrital de servicio a la ciudadanía. Alcaldía Mayor de Bogotá.

Fowler, M. (2002). Patterns of enterprise application architecture. Addison-Wesley.

Guío, A., Velásquez, J., et al. (2021). Marco ético para la inteligencia artificial en Colombia. Ministerio TIC, Departamento Nacional de Planeación & Banco Interamericano de Desarrollo.

Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., & Pineau, J. (2020). Towards the systematic reporting of the energy and carbon footprints of machine learning. Journal of Machine Learning Research, 21(248), 1–43. http://jmlr.org/papers/v21/20-312.html

Jurafsky, D., & Martin, J. H. (2023). Speech and language processing (3rd ed., draft). Stanford University. https://web.stanford.edu/~jurafsky/slp3/

Meta AI. (2024). Introducing LLaMA 3: Open and efficient foundation models. Meta Research. https://ai.meta.com/llama/

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. https://arxiv.org/abs/1301.3781

Mistral AI. (2023). Mistral 7B: Open-weight large language model. https://mistral.ai/

OCDE. (2020). OECD principles on digital government. OECD Publishing. https://www.oecd.org/gov/digital-government/

OCDE. (2020). OECD principles on artificial intelligence. OECD Publishing. https://oecd.ai/en/ai-principles

OpenAI. (2024). GPT-4 Turbo: Pricing and technical overview. https://openai.com/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Secretaría Distrital de Planeación. (2024). Plan Distrital de Desarrollo “Bogotá Camina Segura 2024–2027”. Alcaldía Mayor de Bogotá D. C.

Secretaría General de la Alcaldía Mayor de Bogotá. (2023). Informe de gestión y servicio a la ciudadanía 2023. Alcaldía Mayor de Bogotá D. C.

Secretaría General de la Alcaldía Mayor de Bogotá. (2024). Informe de transformación digital y canales de atención. Alcaldía Mayor de Bogotá D. C.

Secretaría General de la Alcaldía Mayor de Bogotá. (2025). Plan de acción institucional 2025. Alcaldía Mayor de Bogotá D. C.

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645–3650. https://doi.org/10.18653/v1/P19-1355

Tiangolo, S. (2018). FastAPI: Modern, fast (high-performance), web framework for building APIs with Python. https://fastapi.tiangolo.com/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30. https://arxiv.org/abs/1706.03762

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824–24837. https://arxiv.org/abs/2201.11903

Zhang, A., Zhang, X., & Zhao, J. (2024). Multi-agent orchestration for large language models: A survey. arXiv preprint.