#5 EN INNOVACIÓN

RETO DE CIUDAD 1

DATOS CON HISTORIA

DOCUMENTO FINAL DE LA PROPUESTA

Reto de Ciudad 1: Datos con Historia

1. Información del equipo postulante

Nombre del equipo o razón social: AUMENTAR

Tipo de postulante:

X Persona Jurídica

□ Equipo de Innovadores

□ Grupo de Investigación

Composición del equipo:

Complete la siguiente información para cada integrante del equipo (máximo 5 personas):

Nombre completo	Rol en el equipo	Formación	Experiencia relevante
Dolly Gallego	Coordinadora	- Tecnóloga industrial - Ingeniera de producción - Maestría en dirección logística - Diplomado en desarrollo de software	Más de 5 años en gestión de proyectos de inversión pública
Jorge Eduardo Adan Beltrán	Analista de Datos	- Tecnólogo en Sistematización de Datos - Ingeniero en Telemática - Applied Scrum for Project Management	Más de 10 años de experiencia en desarrollo full stack y más de 4 años como arquitecto y administrador de data centers
Christian David Gonzalez Uribe	Ingeniero de Agentes Inteligentes y LLM	- Tecnólogo en electrónica - Ingeniero en control	Desarrollo de aplicaciones de software para entidades

		- Certificado en creación de sistemas multiagentes IA - Google Cloud Vertex AI Agent Builder (Certificado en curso)	públicas como la policía nacional. Desarrollo de firmware y hardware para la multinacional Honeywell. Más de 4 años en el desarrollo de aplicaciones con tecnologías 4.0 (AR e IA)
Jaime Arley Castañeda Useche	Desarrollador Fullstack Mid-level 1	- Tecnólogo en sistematización de datos - Certificado Profesional en Git y GitHub - Certificado en fundamentos de Angular	Más de 2 años como desarrollador web fullstack
Julian Santos Cubides	Desarrollador Fullstack Mid-level 2	- Tecnólogo en Sistematización de Datos - Ingeniería de Sistemas y Computación (En curso)	1 año en el desarrollo web full stack
Julian Enrique López Salinas	Aseguramiento de la calidad (QA)	- Ingeniero mecánico - Diplomado en ISO 9001 - Diplomado en ISO 45001 - Diplomado en ISO 55001	3 años de experiencia en el sector productivo, implementando metodologías de mejora de productividad 2 años de experiencia como tester QA de múltiples plataformas digitales 4 años de experiencia en el sector educativo asegurando el buen funcionamiento en laboratorios 1 año de experiencia en el sector energético

2. Descripción de la solución propuesta

Título de la solución: GIA-Sistema de Gestión de Información Asistida

Resumen ejecutivo:

La Secretaría Distrital de Planeación (SDP) enfrenta un desafío significativo en el manejo de grandes volúmenes de datos cualitativos y cuantitativos provenientes de múltiples fuentes, como mesas de participación ciudadana, actas, documentos de políticas públicas y reportes de proyectos de inversión. La falta de homogeneidad y estructura de estos datos dificulta su análisis, retrasa la integración de observaciones ciudadanas en los procesos de toma de decisiones y limita la capacidad de evaluar el impacto de las políticas y proyectos. Los métodos manuales actualmente empleados generan lentitud, riesgos de error y falta de aprovechamiento de la información..

Nuestra solución tecnológica aborda esta problemática mediante la implementación de un LLM (Large Language Model) ajustado al contexto específico de la SDP. Este modelo automatiza el análisis, categorización e identificación de tendencias y patrones clave en la información cualitativa y cuantitativa. Un sistema de consolidación centraliza los datos de aportes ciudadanos, incluyendo aquellos recopilados a través de la herramienta "Chatico", matrices de seguimiento de políticas públicas y proyectos de inversión, así como documentos base de conocimiento como CONPES, POT y PDD, para garantizar un acceso unificado y continuo. Además, un módulo de limpieza y preparación de datos asegura la precisión y anonimización de la información procesada.

La solución incluye una plataforma web intuitiva con dashboards que permiten a los funcionarios visualizar categorías, tendencias y métricas clave, e interactuar con un chatbot que resuelva consultas sobre los datos analizados y sobre el uso de la plataforma, mejorando las actividades alrededor de la planificación urbana y la toma de decisiones,. Este enfoque proporciona mayor agilidad y precisión en el tratamiento de datos cualitativos y cuantitativos, facilitando la integración efectiva de las observaciones y necesidades ciudadanas en políticas públicas y proyectos de inversión optimizando así los procesos de planeación y participación.

Descripción detallada de la solución:

1. Problema

La problemática central identificada en el Reto de Ciudad 1 se centra en la dificultad para gestionar de manera eficiente la gran cantidad de información cualitativa y cuantitativa manejada por las dependencias de la Secretaría Distrital de Planeación (SDP): Oficina de Participación y Diálogo

Ciudadano (OPDC) , Subsecretaría de Planeación de la Inversión (SPI) y Subsecretaría de Políticas Públicas y Planeación Social y Económica (SPPPSE). Esta situación involucra diversos factores interrelacionados que impactan tanto la calidad de las decisiones estratégicas como la capacidad de responder eficazmente a las necesidades ciudadanas. A continuación, se describen los principales elementos de esta problemática:

El volumen y la desestructuración de los datos cualitativos y cuantitativos es uno de los desafíos clave. Las dependencias recopilan extensas cantidades de información proveniente de múltiples fuentes, como mesas de participación ciudadana, consultas públicas, actas y reportes narrativos. Este tipo de datos suelen estar desestructurados y carecen de un formato homogéneo, lo que dificulta su análisis. Esta falta de estructura retrasa la extracción de conclusiones significativas y, en consecuencia, la integración de aportes ciudadanos y observaciones en las políticas y planes de acción.

El procesamiento manual de estos datos, la interpretación y transformación de la misma para la generación de insumos o reportes, puede tardar semanas, lo que ralentiza los tiempos de respuesta e introduce riesgos de error humano e inconsistencia. La falta de automatización limita el análisis en tiempo real de la información restringiendo su aprovechamiento, por lo cual, los aportes ciudadanos y las observaciones recogidas no se integran de manera efectiva en los procesos de toma de decisiones.. Esto puede ocasionar que información valiosa pase desapercibida o se sub-utilice, afectando la capacidad de que las decisiones de planificación urbana estén alineadas con la realidad y las necesidades de los ciudadanos.

De igual manera, la integración limitada de observaciones sobre avances, problemas o logros impacta negativamente el ajuste de la planificación, la detección oportuna de ineficiencias o retrasos y el monitoreo de proyectos de políticas públicas y de inversión

Esta problemática evidencia la necesidad de automatizar la recolección, análisis y categorización de la información cualitativa gestionada por las distintas dependencias de la SDP mediante el desarrollo de una solución tecnológica que permita integrar efectivamente esta información en la toma de decisiones, optimizando así la planificación urbana y mejorando el impacto de las políticas y proyectos en beneficio de los ciudadanos.

2. Objetivos de la solución

2.1. Objetivo General

Automatizar el proceso de recolección, análisis y categorización de la información cualitativa y cuantitativa generada por la Secretaría Distrital de Planeación mediante una solución tecnológica

basada en LLMs (Modelos de Lenguaje de Gran Escala), con el fin de mejorar la eficiencia en la toma de decisiones y optimizar la planificación urbana.

2.2. Objetivos Específicos

- *Desarrollar un sistema de consolidación de datos que permita integrar y estructurar la información cualitativa y cuantitativa proveniente de diversas fuentes y formatos de la Secretaría Distrital de Planeación, asegurando su limpieza, anonimización y preparación para el análisis automatizado.*

- *Ajustar y entrenar un modelo de lenguaje (LLM) personalizado en español con información específica de la Secretaría Distrital de Planeación (SDP) para procesar grandes volúmenes de datos cualitativos y cuantitativos, permitiendo su análisis y clasificación en función de tipologías y criterios definidos en conjunto con los profesionales de la entidad.*

- *Desarrollar e implementar un sistema automatizado de categorización que identifique, clasifique y etiquete aportes ciudadanos, observaciones e información cualitativa según su tipología y temática. Asimismo, deberá categorizar el nivel de avance de los productos de políticas públicas y proyectos de inversión.*

- *Identificar tendencias y generar insights clave mediante el análisis de los datos procesados por el LLM, proporcionando información valiosa sobre intereses ciudadanos, áreas de mejora, factores comunes en los proyectos de inversión y políticas públicas, y otras métricas relevantes que mejoren la toma de decisiones estratégicas.*

- *Diseñar e implementar una plataforma web que permita a los usuarios cargar documentos y visualizar los resultados de categorización y análisis en tiempo real, ofreciendo dashboards intuitivos con métricas, tendencias y herramientas de consulta personalizadas, mejorando la accesibilidad y comprensión de la información para tomadores de decisiones.*

- *Desarrollar e implementar un chatbot que permita a los funcionarios de la SDP realizar consultas precisas sobre los datos analizados, acceder a información específica relacionada con las Políticas Públicas, el Plan Distrital de Desarrollo, los Objetivos de Desarrollo Sostenible, así como consultar el manual de uso y las buenas prácticas de la plataforma.*

- *Capacitar a los funcionarios de la SDP en el uso de la solución tecnológica, garantizando una adopción efectiva, promoviendo su uso en procesos diarios, y fomentando una cultura de mejora continua basada en la innovación tecnológica y la toma de decisiones informadas.*

3. Metodología propuesta

3.1. Análisis y levantamiento de requerimientos

En esta fase, se identificarán los requerimientos funcionales y no funcionales de la solución tecnológica mediante reuniones y talleres con las diferentes dependencias de la Secretaría Distrital de Planeación para entender más a detalle sus flujos de trabajo actuales y las necesidades específicas de cada área: Oficina de participación y diálogo ciudadano (OPDC), Subsecretaría de políticas públicas (SPPPSE) y Subsecretaría de planeación de la inversión (SPI). Esto incluirá un levantamiento de información sobre las categorías de análisis requeridas, las fuentes de datos cualitativos y cuantitativos, los criterios de anonimización de datos sensibles, y métricas de desempeño.

3.2. Preparación y consolidación de datos

La información cualitativa y cuantitativa existente será recopilada y consolidada en un formato estructurado y estandarizado. Esto incluirá matrices de seguimiento de proyectos, actas de mesas de participación y/o matrices de aportes ciudadanos . Los datos se limpiarán para eliminar inconsistencias y garantizar que sean adecuados para ser procesados por el modelo de lenguaje., y la información sensible será anonimizada para dar cumplimiento a las leyes de protección de habeas data.

3.3. Personalización y ajuste del LLM

El modelo de lenguaje será ajustado utilizando los datos cuantitativos y cualitativos específicos proporcionados por la SDP. Se ajustará el modelo con un enfoque en la comprensión de la semántica y contexto propio de la Secretaría Distrital de Planeación. El ajuste fino garantizará que el LLM sea capaz de categorizar, analizar y detectar patrones y tendencias relevantes conforme a los criterios definidos en conjunto con el personal de la SDP.

Adicionalmente, este ajuste permitirá la implementación de un chatbot que posibilite a los funcionarios de la SDP realizar consultas sobre los datos analizados, permitiendo filtrar información, analizar tendencias específicas o realizar búsqueda de información relevante de los aportes

ciudadanos o el seguimiento de proyectos de inversión y productos de políticas públicas, así como consultas sobre el uso propio de la plataforma.

3.4. Implementación de herramientas de categorización y análisis

En esta etapa, se desarrollarán algoritmos para la categorización o etiquetado automático de información cualitativa según tipologías temáticas y/o niveles de prioridad. Esto incluirá funcionalidades como la identificación de patrones temporales, la detección de temas recurrentes, datos geográficos, correlación de temáticas, generación de alertas y análisis de barreras en proyectos.

3.4. Desarrollo de la plataforma web

Se diseñará e implementará una plataforma web intuitiva que permitirá a los usuarios de la SDP interactuar y emplear las diferentes funciones del sistema. La interfaz incluirá módulos para la gestión de usuarios, acceso seguro según roles y permisos específicos, y carga de archivos como propuestas ciudadanas, actas de participación, seguimiento de políticas públicas y registros de proyectos de inversión. Además, también se podrán visualizar dashboards con los resultados de categorización, análisis de tendencias y métricas clave que faciliten la toma de decisiones estratégicas.

De igual manera, contará con una interfaz de usuario que permita la interacción entre los funcionarios de la SDP y un chatbot que permita realizar consultas sobre los datos analizados y el uso propio de la plataforma.

La información visualizada en los dashboards así como la suministrada por el chatbot se utilizará para la generación de insumos (texto, gráficas y tablas) para la elaboración de reportes.

3.5. Pruebas piloto y validación

Se llevará a cabo un piloto con un subconjunto de datos de cada dependencia para validar el funcionamiento del LLM y la plataforma web. Se trabajará de manera conjunta con los usuarios finales para evaluar la precisión de las categorizaciones, el análisis de tendencias y la efectividad de la interfaz. Se recolectará retroalimentación para realizar ajustes y mejoras antes de la implementación completa; mediante la recolección de logs y la retroalimentación de los funcionarios de la SDP, se itera de manera continua durante esta etapa para la mejora del desempeño del sistema.

3.6. Capacitación y socialización

Se desarrollarán talleres y capacitaciones dirigidos a los funcionarios de la SDP, enseñando el uso de la plataforma y las capacidades de la misma. La capacitación incluirá manuales de usuario, guías de buenas prácticas y soporte técnico inicial. Este proceso asegurará que los funcionarios puedan explotar al máximo las capacidades de la herramienta, incorporando el análisis automatizado en sus procesos de toma de decisiones.

3.7. Mantenimiento y garantía

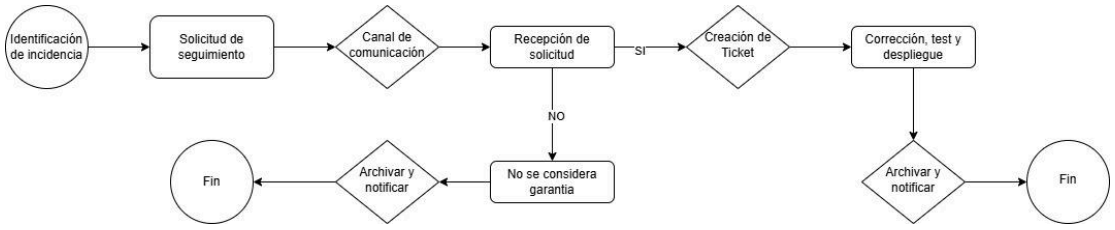
Tras la finalización del contrato y la entrega del proyecto a satisfacción, se ofrecerá un plan de mantenimiento técnico y actualizaciones periódicas durante un período de seis (6) meses. Este plan está diseñado para garantizar el correcto funcionamiento de las funcionalidades desarrolladas, tal como se describen en esta propuesta y las que sean acordadas específicamente con Maloka.

El mantenimiento correctivo por garantía se limitará exclusivamente a las funcionalidades descritas y aprobadas en el alcance del proyecto. Cualquier solicitud relacionada con aspectos no contemplados o funcionalidades adicionales será considerada como un nuevo desarrollo, fuera del alcance de la garantía.

Para garantizar una gestión adecuada de las solicitudes, se implementará un proceso de validación de incidencias. Este proceso permitirá determinar si una solicitud de soporte corresponde a un mantenimiento correctivo bajo garantía o si se trata de un requerimiento que implica un nuevo desarrollo.

Se establecerá un canal de soporte dedicado para atender dudas y resolver incidentes que puedan surgir durante el uso de la solución. El canal principal de comunicación asignado será el correo electrónico ***soportesdp@aumentar.com.co****, a través del cual los usuarios podrán reportar cualquier inconveniente o consulta técnica.*

A continuación se visualiza el diagrama que muestra el flujo de soporte y seguimiento desde la incidencia hasta la solución.



4. Componentes principales de la solución

4.1. LLM personalizado en español:

El LLM será configurado para procesar y analizar grandes volúmenes de datos cualitativos y cuantitativos proporcionados por la SDP, tales como aportes ciudadanos, actas de participación, matrices de seguimiento de políticas públicas y proyectos de inversión. Este modelo será ajustado específicamente para comprender la semántica y el contexto propios de la SDP, integrando terminología y referencias clave de procesos de planificación urbana y documentos base de conocimiento como el CONPES, POT, ODS y PDD. Esto permitirá a los funcionarios consultar información relevante y utilizarla para el análisis cualitativo, la toma de decisiones y la generación de reportes.

El llm se ajustará con dos propósitos:

1. *Análisis de datos y categorización automática:*

   Matriz de productos de políticas públicas: Se analizará información cualitativa de cada producto, combinándola con indicadores cuantitativos para determinar el estado de avance de su implementación. Los productos se clasificarán en tres niveles:

   - *Rojo: Hay retrasos o falta de información para determinar el avance.*
   - *Amarillo: El indicador está en progreso.*
   - *Verde: El producto se ha completado.*

     Se generarán alertas si la descripción cualitativa es incoherente con el indicador.

   Por otra parte, se analizará la información cualitativa y de enfoque de cada producto para identificar y etiquetar los reportes con las siguientes variables estratégicas : Acción climática, Movilidad, Pobreza, Calidad de educación, Acceso a Salud y Seguridad.

   Matriz de proyectos de inversión: Se procesará la información cualitativa y cuantitativa para evaluar el avance hacia las metas del PDD. Se aplicarán las siguientes categorías:

   - *Rojo: Hay retrasos o falta de información para determinar el avance.*
   - *Amarillo: El indicador está en progreso.*
   - *Verde: Se ha alcanzado la meta.*

   Se generarán alertas si la descripción cualitativa es incoherente con el indicador.

De igual manera se realiza un análisis para identificar áreas temáticas desarrolladas en el proyecto tales como vivienda, salud, seguridad, mujer, movilidad, medio ambiente o educación.

Matrices de aportes ciudadanos, aportes ciudadanos hechos a través del chatbot Chatico y actas de mesas de participación ciudadana: El LLM realizará un análisis de sentimiento para identificar los temas de interés de los ciudadanos y etiquetará cada aporte con ejes temáticos como vivienda, salud, seguridad, mujer, movilidad, medio ambiente, educación, entre otros según se acuerde con la SDP. De igual manera se etiqueta el aporte como Comentario, Queja, Propuesta o Recomendación.

Por último, el LLM realizará un análisis inicial de patrones y tendencias clave como causas recurrentes en el atraso del cumplimiento de metas de proyectos de inversión, cuáles políticas públicas presentan menor o mayor avance, correlación entre sectores líderes con otros sectores, necesidades más importantes para la ciudadanía, relación entre necesidades de la ciudadanía e información geográfica, relación entre ejes temáticos de participación ciudadana, entre otros, según se acuerde con los funcionarios de las tres dependencias de de la entidad.

2. *Chatbot: Implementación de un chatbot que posibilite a los funcionarios de la SDP realizar consultas sobre los datos analizados, permitiendo filtrar información, analizar tendencias específicas o realizar búsqueda de información relevante de los aportes ciudadanos o el seguimiento de proyectos de inversión y productos de políticas públicas, así como consultas sobre el uso propio de la plataforma.*

   Este chatbot consta principalmente de dos componentes:

   - *Query Filter: Este se encarga de analizar el mensaje recibido del usuario y determinará si el sistema puede resolver o no la solicitud recibida. En los casos en que el mensaje no se entienda o el mensaje contenga una solicitud que se salga del alcance de las funcionalidades del sistema, la solicitud será clasificada como "Solicitud no Identificada". En caso contrario la solicitud será clasificada en una de dos categorías "Análisis de Información" y "Manual de Uso". Este filtro se aplica para evitar que el LLM intente resolver problemas para los cuales no fue entrenado o ajustado y evitar errores en el sistema o respuestas erróneas.*

   - *Query Handler: Se encarga de dar cumplimiento a la solicitud del usuario. A partir de la información procesada, el LLM podrá responder preguntas específicas sobre la información de las matrices de seguimiento de políticas públicas. proyectos de inversión, o aportes ciudadanos, tales como estado de avance acumulado por*

política pública, objetivos del PDD con mayor número de metas cumplidas, temas de mayor interés de la ciudadanía o qué problemática aqueja más a una localidad en específico. El alcance de este tipo de solicitudes que podrá resolver el chatbot se definirá con los funcionarios de las tres dependencias de la SDP según sus necesidades y prioridades, y se establecerá un número limitado de este tipo de solicitudes según su complejidad.

Adicionalmente, el usuario podrá consultar los documentos base de conocimiento tales como CONPES, POT y PDD y extraer información relevante o resúmenes, con el objetivo de apoyar el análisis e identificación de cumplimiento de las metas del PDD o los aportes de políticas públicas a los objetivos del PDD y ODS.

Asimismo, chatbot estará en la capacidad de resolver dudas que conciernen al manual de usuario de la aplicación y las buenas prácticas de uso de la misma.

4.2 Consolidación de Información:

Este módulo centralizará los documentos base de conocimiento como CONPES, POT y PDD así como las matrices de seguimiento de productos de políticas públicas, matrices de seguimiento de proyectos de inversión, matrices de aportes ciudadanos, información recopilada a través del chatbot 'Chatico' y actas de mesas de participación ciudadana que la SDP desee analizar. La integración se realizará de manera automatizada, consultando documentos alojados en la nube mediante SharePoint.

4.3 Limpieza y Preparación de Datos

Procesará la información consolidada para asegurar condiciones óptimas para el análisis. Incluye tareas de eliminación de datos duplicados, corrección de errores y la segmentación de textos largos en unidades manejables. También se aplicará un proceso de anonimización de datos sensibles y una segmentación basada en criterios definidos por la SDP.

4.4. Plataforma web para carga y visualización:

La solución incluirá una plataforma web que permitirá a los funcionarios consultar documentos de manera sencilla, gestionar roles y permisos y visualizar los resultados obtenidos del análisis de datos del LLM a través de dashboards interactivos. Estos dashboards mostrarán análisis de tendencias, categorías, métricas clave y resúmenes ejecutivos que facilitarán la toma de decisiones estratégicas.

Adicionalmente, esta plataforma facilitará la generación de insumos (texto, gráficas y tablas) para la elaboración de reportes basados en la información visualizada en los dashboards y la información suministrada por el chatbot.

4.5. Seguridad y manejo de datos:

La solución asegurará la anonimización y manejo seguro de cualquier información sensible, siguiendo las mejores prácticas de seguridad. El acceso a la plataforma será gestionado por roles de usuario, permitiendo que diferentes niveles de funcionarios tengan acceso adecuado según sus responsabilidades.

5. Resultados esperados

La implementación de la solución propuesta permitirá transformar el manejo de la información cualitativa en la Secretaría Distrital de Planeación (SDP), generando resultados significativos en términos de eficiencia, precisión y toma de decisiones informadas. Los principales resultados esperados son:

- *Reducción en el tiempo de análisis de la información cualitativa: La automatización del procesamiento de datos cualitativos disminuirá los tiempos asociados a la recolección, clasificación y análisis de información que actualmente se realizan de forma manual. Con esta mejora, se logrará responder de manera más ágil y eficiente a las necesidades ciudadanas, permitiendo una toma de decisiones más rápida y mejor fundamentada.*

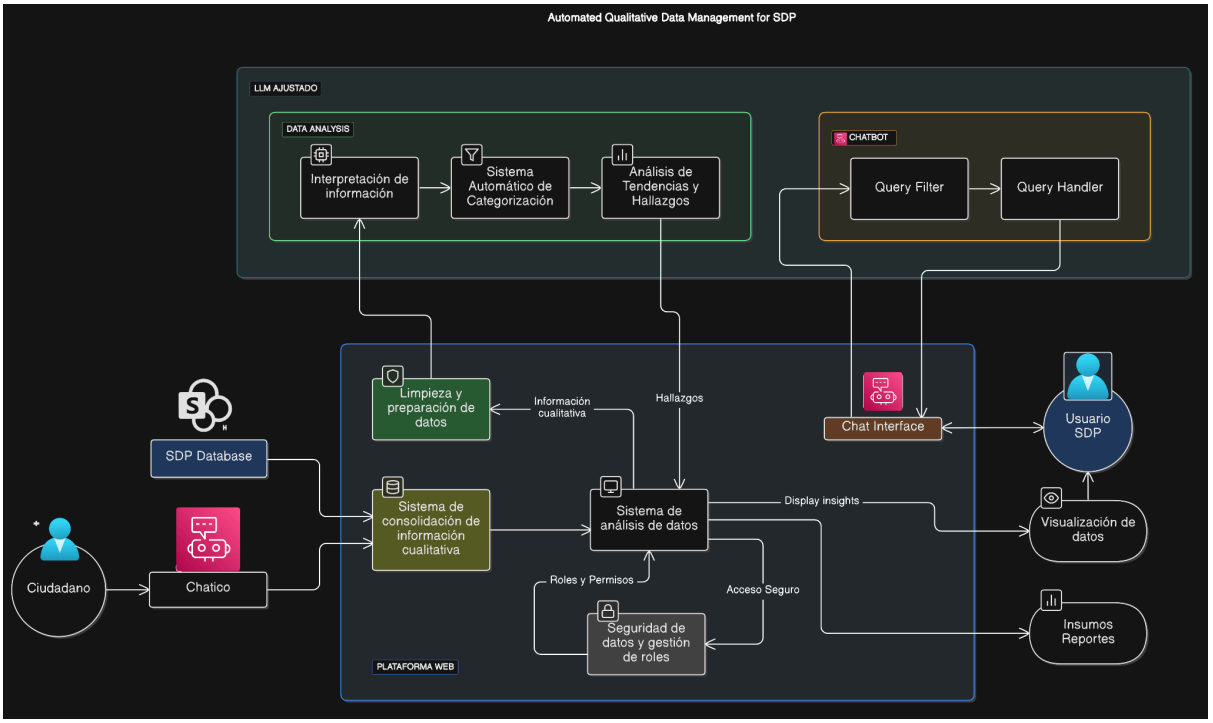
- *Categorización de información cualitativa: La solución permitirá clasificar de manera automática y consistente la información cualitativa según tipologías temáticas, niveles de prioridad y estados de avance. Esto garantizará un manejo más ordenado y accesible de los datos, facilitando su análisis y uso por parte de los funcionarios.*

- *Identificación de tendencias y patrones clave: El análisis automatizado permitirá detectar de manera proactiva patrones y tendencias relevantes dentro de la información cualitativa. Esto incluirá la identificación de demandas ciudadanas recurrentes, evaluación de la efectividad de políticas públicas y monitoreo de áreas de mejora en los proyectos de inversión.*

- *Transformación e Integración de la información procesada y categorizada: Implementación de dashboards interactivos y accesibles para los funcionarios, lo que facilitará la visualización de datos clave. De igual manera, el desarrollo de un chatbot que permita realizar consultas sobre los datos analizados, extraer o resumir información específica de los documentos base de Políticas Públicas, Plan Distrital de Desarrollo y Objetivos de Desarrollo Sostenible, así como consultas sobre el uso y buenas prácticas de la plataforma. Esto permitirá a la SDP tomar decisiones más informadas y estratégicas.*

3. Arquitectura de la solución

Diagrama de Flujo:



Descripción de componentes:

La arquitectura de la solución combina dos componentes clave para garantizar un flujo eficiente de procesamiento y análisis de datos. En primer lugar, se emplearán tecnologías web avanzadas para desarrollar una plataforma interactiva que permita a los funcionarios de la SDP cargar documentos, acceder a dashboards y visualizar resultados mediante herramientas intuitivas. En segundo lugar, se integra un LLM ajustado específicamente a las necesidades de la SDP, que automatiza el análisis y la categorización de la información cualitativa, detectando tendencias y patrones relevantes.

El sistema de administración *es el encargado de coordinar los demás módulos y hacer que estos funcionen al unísono. Es el núcleo de la plataforma web para la cual se emplearían los siguientes framework: Laravel, VueJS, Streamlit y Python.*

El sistema de consolidación de información cualitativa *tiene como objetivo centralizar los documentos base de conocimiento como CONPES, POT y PDD así como las matrices de seguimiento de políticas públicas,*

de proyectos de inversión y de aportes ciudadanos, información recopilada a través del chatbot 'Chatico' y actas de mesas de participación ciudadana que la SDP desee analizar. La integración se realizará de manera automatizada, consultando documentos alojados en la nube mediante SharePoint.

Los datos estructurados (matrices y tablas) se integran directamente en el sistema, mientras que los no estructurados (Archivos PDF, transcripciones o documentos digitalizados) pasarán primero por un pre procesamiento, mediante técnicas OCR (Reconocimiento Óptico de Caracteres) para extraer su contenido y dar un formato a la información para poder integrarlo al sistema. Cabe aclarar que los archivos PDF o documentos digitalizados deben cumplir un mínimo de calidad para poder procesarlos posteriormente de manera efectiva.

Para el correcto funcionamiento de este módulo, es indispensable contar con los permisos y accesos necesarios a las bases de datos y APIs de la SDP.

***El componente de seguridad de datos y gestión de roles** se encarga de garantizar que todas las conexiones a bases de datos o APIs de la SDP se realicen de manera segura, utilizando protocolos de cifrado HTTPS y AOuth para proteger la integridad de la información durante la conexión. Asimismo, este módulo protege el acceso al sistema mediante un proceso de inicio de sesión seguro, validando la identidad de los usuarios autorizados. Adicionalmente, gestiona los permisos asociados a los diferentes roles de usuarios, limitando o habilitando el acceso a funcionalidades específicas según su nivel de autorización.*

***El sistema de limpieza y tratamiento de datos** procesa la información recopilada por el sistema de consolidación para prepararla antes de ser analizada por el LLM. Este módulo se encarga de eliminar datos duplicados, corregir errores y segmentar textos extensos en unidades manejables, garantizando consistencia y precisión en su posterior análisis. Además, se aplicarán procesos de anonimización para proteger la información sensible y asegurar el cumplimiento de las leyes de protección de datos. El objetivo es transformar datos desestructurados en entradas organizadas y coherentes, listas para ser procesadas por el LLM, maximizando la calidad y relevancia del análisis.*

*El módulo encargado de realizar el análisis de los datos será el **LLM Ajustado.** Primero, el LLM procesa los datos recibidos del módulo de limpieza y tratamiento de datos, utilizando técnicas de procesamiento de lenguaje natural. Luego, aplica una categorización o etiquetado de los datos en categorías predefinidas por la SDP. Posteriormente, realiza un análisis de tendencias para identificar patrones y generar hallazgos clave.*

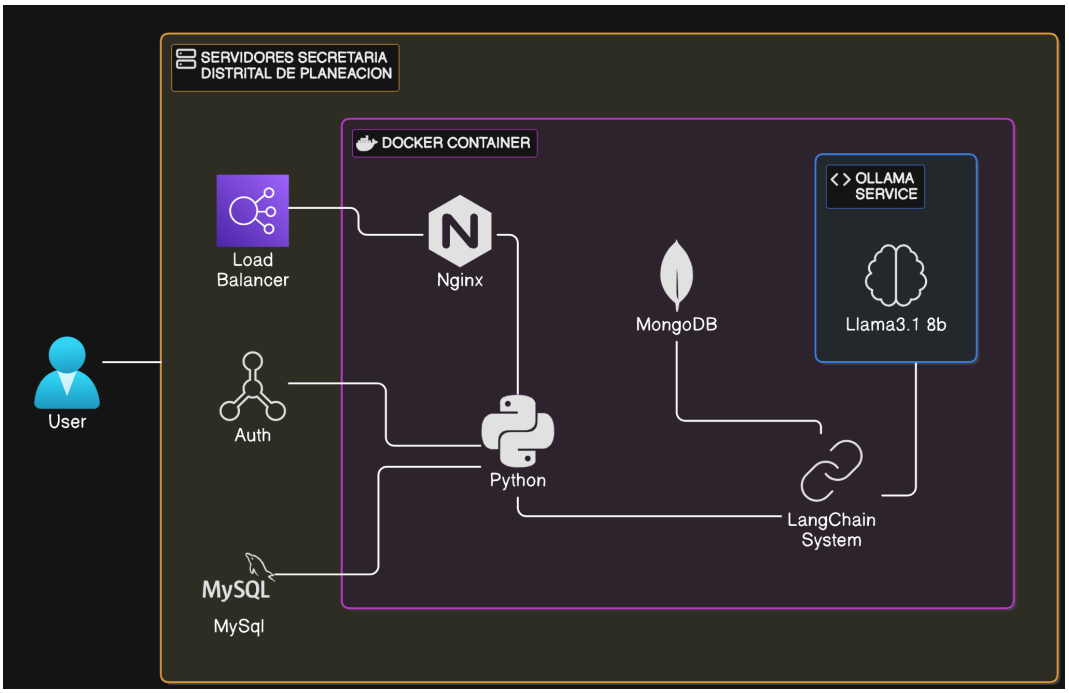
*Al mismo tiempo, el LLM ajustado analiza, filtra y resuelve las consultas realizadas por los funcionarios de la SDP, en relación a los datos analizados por el LLM y el manual de uso y buenas prácticas de la aplicación, a través del **chatbot interface** de la plataforma.*

*Este LLM estará basado en el modelo **Llama 3.1,** una solución open source que se ajustará mediante el uso de Python y el framework Langchain. Esta elección, junto con el uso de la aplicación Ollama, permite la instalación local del modelo en los servidores de la SDP, asegurando que toda la información se mantenga dentro del entorno seguro y controlado de la entidad, sin la necesidad de recurrir a APIs de terceros, lo que garantiza la integridad y confidencialidad de los datos procesados.*

***El componente de visualización** de resultados proporciona una interfaz de usuario en donde los funcionarios de la SDP pueden visualizar los análisis y conclusiones generadas por el modelo LLM. Los usuarios podrán consultar categorizaciones, tendencias, métricas clave y hallazgos relevantes de manera clara y accesible a través de dashboards.*

Diagrama de Arquitectura

Alternativa 1 - Entorno Local



La arquitectura de la aplicación se encuentra organizada en contenedores Docker y algunos de servicios:

*El balanceador de carga (**Load Balancer**), que distribuye el tráfico.*

*Autenticación (**Auth**) gestiona la seguridad y los accesos.*

*El backend utiliza un contenedor **Docker** con:*

- ***Nginx** como servidor web.*

Python *para la lógica de negocio y conexión con sistemas externos.*

MongoDB *como base de datos NoSQL para almacenar documentos o datos no estructurados.*

MySQL *como base de datos relacional.*

LangChain System *para gestionar tareas complejas y uso de herramientas para integrar e interactuar con el LLM.*

Ollama Service *Interpreter que permite la ejecución del modelo de lenguaje* ***Llama 3.1-8B.***

Esta propuesta busca implementar una solución tecnológica ejecutando el modelo de lenguaje LLM Llama 3.1-8B de manera local dentro de los servidores de la Secretaría Distrital de Planeación (SDP). Esta configuración tiene como objetivo principal garantizar que la información sensible y estratégica de la SDP permanezca privada y segura, eliminando la necesidad de exponer datos a servicios o API externas.

Al ejecutar el LLM localmente, se evitan riesgos asociados con la transferencia de datos sensibles a terceros, además de cumplir con normativas estrictas de manejo de datos. Otra ventaja significativa es la mejora en los tiempos de respuesta: al no depender de una conexión externa, se elimina la latencia que podría generarse al enviar y recibir datos desde APIs externas, lo cual es crítico para manejar altos volúmenes de solicitudes simultáneas. Asimismo, esta alternativa reduce los costos asociados al tráfico externo, ya que las llamadas a APIs de terceros suelen tener un costo variable que aumenta con el uso intensivo.

Los requisitos mínimos de hardware para la ejecución de este modelo de lenguaje se describen a continuación:

CPU: Intel Core i7-9700K or AMD Ryzen 9 5900X (o equivalente)

GPU: 16 GB of VRAM

RAM: 64 GB DDR4 RAM

Storage: 1 TB NVMe SSD

Sistema Operativo: 64-bit Linux

Los requisitos recomendados de hardware para la ejecución de este modelo son:

CPU: Intel Core i9-11900K or AMD Ryzen Threadripper 3960X (o equivalente)

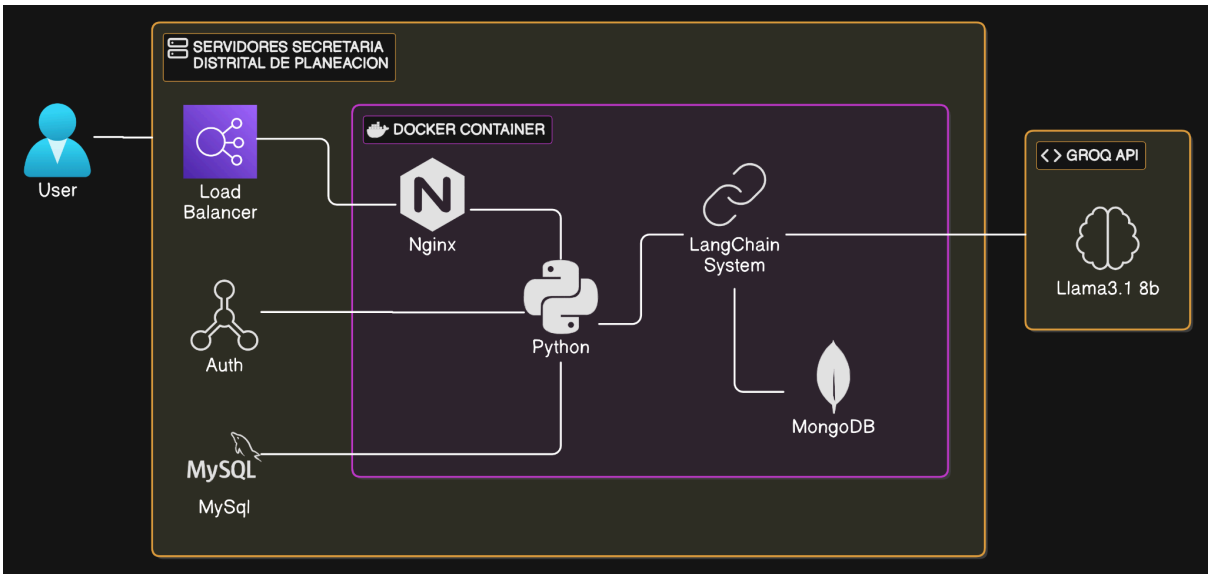
GPU: 24 GB of VRAM

RAM: 128 GB DDR4 RAM

Storage: 2 TB NVMe SSD

Sistema Operativo: 64-bit Linux

Alternativa 2 - API Externa



*El balanceador de carga (**Load Balancer**), que distribuye el tráfico.*

*Autenticación (**Auth**) gestiona la seguridad y los accesos.*

*El backend utiliza un contenedor **Docker** con:*

- ***Nginx** como servidor web.*
- ***Python** para la lógica de negocio y conexión con sistemas externos.*
- ***MongoDB** como base de datos NoSQL para almacenar documentos o datos no estructurados.*
- ***MySQL** como base de datos relacional.*
- ***LangChain System** para gestionar tareas complejas y uso de herramientas para integrar e interactuar con el LLM.*
- ***GROQ API** Interpreter que permite la ejecución del modelo de lenguaje **Llama 3.1-8B.***

Como alternativa, se propone utilizar una API externa para la ejecución del modelo de lenguaje LLM, considerando las posibles limitaciones de los recursos de hardware en los servidores de la Secretaría Distrital de Planeación (SDP). Esta opción permite acceder a modelos avanzados sin necesidad de realizar inversiones significativas en infraestructura local, garantizando un rendimiento óptimo y escalabilidad para procesar grandes volúmenes de datos. Aunque esta opción requiere transferir cierta información a terceros, se pueden implementar medidas adicionales de cifrado y anonimización para proteger la privacidad y cumplir con las normativas vigentes.

Esta alternativa, sin embargo, presenta una limitación importante: las APIs externas en sus versiones gratuitas suelen imponer restricciones en la cantidad de peticiones diarias permitidas. Dado que una de las restricciones clave del proyecto es evitar el pago de licencias, esta opción podría limitar el desempeño de la aplicación en su versión gratuita en situaciones de análisis con alta demanda.

En general, estas arquitecturas modulares, combinadas con tecnologías y herramientas open-source, permiten una alta escalabilidad, ya que facilita la incorporación de nuevos módulos y la adaptación a cambios futuros sin necesidad de reestructuraciones significativas.

Tecnologías propuestas:

Frameworks Web:

- ☐ *Django*
- ☐ *Angular*
- ☐ *React*
- ☐ *Node.js*
- ☐ *Otro (especifique): Laravel, VueJS, Streamlit, Langchain, Ollama*

Lenguajes de Programación:

- ☐ *Java*
- X *Python*
- ☐ *R*
- ☐ *Otro (especifique): PHP, Javascript*

Bases de Datos:

- ☐ *PostgreSQL*
- X *MySQL*
- ☐ *Oracle*
- X *Otro (especifique): MongoDB*

Modelos de IA/NLP:

- ☐ *BERT*
- ☐ *RoBERTa*
- ☐ *XLNet*

□ *T5*
□ *mBART*
□ *No se emplean modelos NLP ni LLMs*
X *Otro (especifique):*
- *Llama 3.1 - 8B*

4. Sobre los datos

Estrategia de análisis de datos:

La metodología propuesta para el procesamiento de los datos estructurados y no estructurados para su posterior análisis posee un enfoque secuencial con el fin de garantizar una extracción de la información precisa.

En primera instancia, los datos estructurados, como tablas y matrices, serán integrados al sistema directamente mediante el sistema de consolidación de información cualitativa. Por otro lado, los datos no estructurados (archivos pdf, transcripciones de texto o documentos digitalizados) serán convertidos a formatos estructurados utilizando métodos de OCR (Reconocimiento Óptico de Caracteres) para extraer el contenido en un formato comprensible para el sistema.

Posteriormente, estos datos serán enviados al sistema de limpieza y tratamiento de datos. Aquí, primero se eliminan duplicados y errores tipográficos para después implementar procesos de anonimización para la protección de datos sensibles. Por último, para los textos más extensos, se aplicarán técnicas de fragmentación y etiquetado para un manejo y posterior análisis más preciso y para mejorar la capacidad de procesamiento puesto que los modelos de lenguaje poseen límites en la cantidad de texto que pueden procesar en una sola entrada.

Para la categorización automática, se llevará a cabo el ajuste fino del LLM utilizando un muestreo de datos específicos proporcionados por la SDP para que el modelo comprenda el contexto y semántica empleada en los procesos de la entidad. Este proceso se complementa con técnicas como la frecuencia de palabras, análisis de sentimientos, análisis de temas y reconocimiento de patrones dentro de los textos, para permitir al LLM reconocer patrones temáticos ("vivienda", "salud", "transporte", "seguridad", etc), tipologías específicas relevantes para las dependencias de la SDP, como "quejas", "propuestas" o "acuerdos", y hacer análisis de la información cualitativa y cuantitativa de los documentos de control de proyectos de políticas públicas y de inversión, posibilitando la identificación y etiquetado del estado de los proyectos (en curso , bloqueado, a tiempo o en retraso).

La aplicación de estos métodos junto con el uso de algoritmos estadísticos facilitará la detección de patrones de comportamiento y temas recurrentes en la participación ciudadana, permitiendo a la SDP enfocar esfuerzos en áreas de mayor interés o problemática, así como razones comunes que afectan el avance de los proyectos de políticas públicas y de inversión y predicciones o identificación de patrones que permitan generar alertas tempranas o ineficiencias.

Fuentes de datos a utilizar:

Para realizar el ajuste fino del modelo y para la implementación de la solución se necesitaría los siguientes conjuntos de datos:

- *Matriz de aportes de participación ciudadana*
- *Actas de espacios de participación ciudadana*
- *Matriz de propuestas ciudadanas recibidas a través del chatbot 'Chatico'*
- *Matriz de seguimiento de políticas públicas*
- *Matriz de seguimiento de proyectos de inversión*

5. Principales funcionalidades

Listado y descripción de las funcionalidades:

Funcionalidad	*Descripción*	*Usuarios objetivo*	*Valor agregado*
Carga automática de datos	*Carga automática de documentos provenientes de bases de datos de la SDP, así como del chatbot Chatico.*	*No Aplica*	*Se elimina la necesidad de intervención manual por parte de funcionarios de la SDP*
Administración de usuarios	*Se pueden crear, editar y borrar usuarios del sistema, asignar credenciales de inicio de sesión y gestionar distintos niveles de acceso*	*- Funcionario SDP (Administrador)*	*Es un mecanismo que permite definir quiénes tienen acceso a la plataforma*

Login	*Se da acceso a la plataforma a través de un inicio de sesión seguro, limitando o habilitando funciones y vistas del sistema según el nivel de acceso del usuario*	*- Funcionario SDP* *- FuncionarioSPD (Administrador)*	*Proporciona un mecanismo que limita el acceso y uso de la información sólo a personal autorizado*
Visualización de resultados	*Interfaz de usuario en donde se visualizan los análisis y conclusiones generadas por el modelo LLM.*	*- Funcionario SDP* *- FuncionarioSPD (Administrador)*	*A través de dashboards interactivos con gráficos, tablas, opciones de filtrado u otros componentes visuales se pueden realizar análisis más detallados*
Chatbot	*Chatbot que permite realizar consultas sobre los datos analizados por el LLM y sobre el manual de usuario y buenas prácticas de la plataforma*	*- Funcionario SDP* *- Funcionario SDP (Administrado)*	*Se puede obtener información específica de los datos analizados, permitiendo filtrar información, analizar tendencias específicas o realizar búsqueda de información relevante de los aportes ciudadanos o el seguimiento de proyectos de inversión y productos de políticas públicas.* *Además, al manejar la información de manual de usuario y buenas prácticas a través de un chat el usuario obtiene la información específica que necesita a demanda.*
Generación de insumos para reportes	*Se exportan textos ,gráficas y tablas*	*- Funcionario SDP*	*Se obtienen insumos con información estratégica necesarios para la generación de reportes*

6. Innovación y diferenciación

Componente innovador:

Lo que distingue a esta solución es la combinación de un LLM ajustado al contexto específico de la secretaría Distrital de Planeación y la integración de múltiples fuentes de datos cualitativos de manera automática, permitiendo a la entidad transformar en poco tiempo grandes volúmenes de información no estructurada en datos organizados, procesados y valiosos para la toma de decisiones.

La integración del LLM con Langchain, un framework de código abierto robusto para la construcción de aplicaciones basadas en LLM, permite no solo una categorización automática, sino también un análisis de tendencias y patrones que aportan conclusiones significativas y en tiempo real. Además, la implementación de una interfaz intuitiva y accesible, ofrece a los funcionarios de la SDP herramientas visuales y dinámicas que facilitan la consulta y organización de los resultados así como análisis más detallados.

Este enfoque permite una gestión ágil, segura y eficiente de los datos cualitativos que han sido tradicionalmente difíciles de manejar y analizar, contribuyendo a mejorar la efectividad de los proyectos de políticas públicas e inversión y la integración de la participación ciudadana en los procesos de toma de decisiones.

Por último, cabe mencionar que se transforma la información de PDF's y matrices de seguimiento a dashboards interactivos y accesibles para los funcionarios, lo que facilitará la visualización de datos clave. De igual manera, el desarrollo de un chatbot que permita realizar consultas sobre los datos analizados, extraer o resumir información específica de los documentos base de conocimiento, así como consultas del manual de uso y buenas prácticas rompe el paradigma de capacitaciones e instructivos de uso de aplicaciones convencional mediante PDF's o presentaciones.

Ventaja competitiva:

Con un enfoque centrado en la empatía y el trabajo en equipo, asumimos los problemas y necesidades de nuestros clientes como propios, garantizando soluciones diseñadas a la medida. Nuestra trayectoria incluye ***4 años de experiencia*** *en el desarrollo de soluciones interactivas e innovadoras, complementada por los* ***más de 7 años de experiencia individual*** *de cada integrante del equipo, lo que nos permite ofrecer un enfoque integral y multidisciplinar en diversos sectores y disciplinas. Esta sinergia entre expertise, innovación y sensibilidad hacia las necesidades del cliente asegura nuestra capacidad para entregar resultados excepcionales y únicos en el mercado.*

*Además, lo que diferencia a esta propuesta es el uso del modelo de lenguaje **Llama 3.1**, el cual es el LLM de código abierto más avanzado en el mercado actualmente y soporta múltiples lenguajes incluído el español. Desarrollado y respaldado por Meta, ha permitido que gran parte de la comunidad acoja con rapidez estas nuevas tecnologías para diversos casos de uso, como lo es el de la SDP. Al ser de código abierto no se requiere de ningún pago para su uso, generando reducción en los costos de la implementación de la solución propuesta.*

Adicionalmente, gracias al trabajo de cuantización que ha hecho Meta para compactar sus modelos y eliminar la necesidad de grandes infraestructuras de hardware, y a herramientas libres como Ollama es posible instalar, ajustar y ejecutar localmente el LLM en los servidores de la SDP, asegurando que toda la información se mantenga privada y dentro del entorno seguro y controlado de la entidad, sin necesidad de enviar los datos a APIs externas, como ocurre con otros modelos comerciales, garantizando así la integridad, seguridad y confidencialidad de la información.

7. Madurez de la solución

Nivel de madurez actual:

- □ *TRL 1: Principios básicos observados*
- □ *TRL 2: Formulación del concepto de la tecnología y/o aplicación*
- X *TRL 3: Prueba de concepto*
- □ *TRL 4: Validación en laboratorio*
- □ *TRL 5: Validación en entorno relevante*

Avances y desarrollos previos:

Se han realizado pruebas de concepto alrededor de la instalación, ejecución e implementación de LLMs para el análisis de sentimiento y la categorización de textos. Se ha trabajado en el control del LLM mediante la configuración de prompts y el ajuste fino (fine-tuning) del modelo, con el fin de personalizar su comportamiento y mejorar su desempeño en tareas específicas. Los resultados obtenidos han sido en su mayoría satisfactorios, pudiendo distinguir y categorizar textos conforme a ciertas restricciones y categorías pre establecidas, y asegurando que el LLM entregue respuestas con un formato o estructura específicos para facilitar su posterior uso, análisis e integración con otros sistemas.

8. Oportunidades de mejora para futuras implementaciones

Se ha propuesto una solución con una arquitectura que permite la escalabilidad de la misma lo cual permite además la implementación de nuevas implementaciones. Dentro de ellas, se logran identificar algunos posibles futuros desarrollos:

Generación automática de reportes para la SPPPSE: *Generación automatizada de aportes de los productos a los objetivos definidos en el PDD y los ODS.*

Trazabilidad en tiempo real de los aportes: *Trazabilidad de los aportes ciudadanos durante todo el ciclo del flujo de trabajo de la Oficina de Participación Ciudadana.,*

Feedback de matrices de seguimiento: *Se evidencia que un factor común que dificulta el análisis de la información cualitativa es el mal diligenciamiento de las matrices de seguimiento. Una realimentación que indique a los usuarios que están trabajando sobre los registros generaría alertas tempranas sobre información cualitativa que no tiene coherencia con los productos de políticas públicas, los proyectos de inversión y/o las metas del PDD.*

Chatbot como canal de comunicación entre las dependencias: *Mejor integración y análisis de información entre dependencias, fomentaría una mayor colaboración y trazabilidad entre las áreas de Participación Ciudadana, Políticas Públicas y Proyectos de Inversión. Al centralizar y automatizar el análisis de la información cualitativa, y al ofrecer un canal directo de comunicación, se mejorará la coherencia y la comunicación entre dependencias, permitiendo que los esfuerzos estén mejor coordinados y orientados a lograr resultados significativos y medibles para la ciudadanía.*

9. Documentación adjunta

- *Anexo 1. Cronograma*
- *Anexo 2. Presupuesto*

Firma de los integrantes del equipo

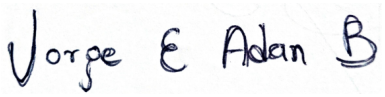


#5 EN INNOVACIÓN

RETO DE CIUDAD 1

DATOS CON HISTORIA

Christian David Gonzalez Uribe
C.C. 1032462368



Jorge Eduardo Adan Beltran
C.C. 1032462603



Dolly Andrea Gallego Narvaez
C.C. 1024543942



Jaime Arley Castañeda Useche
C.C. 1030700020



Julian Santos Cubides
C.C. 1000064026



Julian Enrique López Salinas

#5 EN INNOVACIÓN

RETO DE CIUDAD 1

DATOS CON HISTORIA

C.C. 1016058668