Prototipo de agente conversacional inteligente para el soporte técnico en una empresa de servicios tecnológicos

Lady Stefany Avella Hernández

Director

Isaac Esteban Camargo Freile

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia y Analítica de Datos

2026

Resumen

Este proyecto surge de la necesidad de una empresa de servicios tecnológicos, dedicada a soluciones de facturación electrónica y gestión administrativa. En los últimos años, el aumento de usuarios generó una sobrecarga en el equipo de soporte técnico, especialmente en horarios nocturnos, lo que se traducía en demoras, errores frecuentes y baja satisfacción de los clientes.

Para responder a esta problemática, se diseñó un prototipo de agente conversacional analítico integrado que, a diferencia de un chatbot tradicional, gestiona procesos de consulta sobre la aplicación de forma autónoma. La confiabilidad de las respuestas se garantiza mediante una arquitectura RAG local basada en fuentes y manuales técnicos verificados. La metodología empleada fue CRISP-DM, utilizando un motor de *backend* estructurado para el procesamiento de lenguaje natural (NLP), la normalización de texto y el análisis de patrones de registros de servicio.

La evaluación del sistema en un entorno de producción validó una optimización operativa masiva, logrando reducir el tiempo de primera respuesta de un periodo histórico manual de 2,4 horas a un intervalo de 1,8 a 2,9 segundos por interacción. El prototipo alcanzó una tasa de resolución autónoma global del 78% mediante respuestas FAQ e interacción con herramientas administrativas WDM, derivando únicamente el 22% de casos complejos al canal humano. La prueba de consistencia frente a un Ground Truth de 45 registros demostró una precisión conceptual del 100% en las inferencias, mitigando por completo las alucinaciones comerciales gracias a una adherencia perfecta del 100% a la regla de detección de brechas de conocimiento (*knowledge gaps*), las cuales se capturan y resuelven a través de un circuito cerrado de aprendizaje supervisado. Esta solución disminuye los errores derivados del uso incorrecto del

software de facturación y fortalece la eficiencia de la organización mediante una acceso ágil, seguro y confiable al soporte de la plataforma.

Palabras clave: Agente conversacional inteligente, RAG, Soporte técnico, WDM, *Knowledge Gap*s.

Abstract

This project arose from the needs of a technology services company dedicated to electronic invoicing solutions and administrative management. In recent years, the increase in users generated an overload on the technical support team, especially during nighttime hours, resulting in delays, frequent errors, and low customer satisfaction.

To address this issue, an integrated analytical conversational agent prototype was designed which, unlike a traditional chatbot, autonomously manages application-related query processes. The reliability of the responses is ensured through local RAG-based architecture supported by verified technical sources and manuals. The methodology employed was CRISP-DM, using a structured *backend* engine for natural language processing (NLP), text normalization, and service log pattern analysis.

The system evaluation in a production environment validated massive operational optimization, reducing the first response time from a historical manual average of 2,4 hours to an interval of 1,8 to 2,9 seconds per interaction. The prototype achieved an overall autonomous resolution *Rate Limit* of 78% through FAQ responses and interaction with WDM administrative tools, forwarding only 22% of complex cases to the human support channel. The consistency test against a Ground Truth *dataset* of 45 records demonstrated 100% conceptual accuracy in the inferences, completely mitigating commercial hallucinations thanks to perfect adherence (100%) to the *knowledge gaps* detection rule, which are captured and resolved through a closed-loop supervised learning process. This solution reduces errors derived from the incorrect use of invoicing software and strengthens organizational efficiency through agile, secure, and reliable access to platform support.

Keywords: Intelligent conversational agent, RAG, Technical support, WDM, *Knowledge Gaps.*

Tabla de Contenido

Lista de Figuras

Lista de Apéndices

Introducción

El avance de la digitalización de los procesos administrativos ha llevado a muchas organizaciones a buscar herramientas que faciliten la facturación, el control de los inventarios y la gestión financiera en general. En este escenario, el soporte al cliente se convierte en un componente transversal y fundamental para garantizar que los usuarios comprendan el funcionamiento técnico y normativo de estas plataformas. Tal como mencionan Pahi et al. (2024), la adopción de estrategias tecnológicas en el sector servicios ha demostrado ser un factor determinante para la competitividad, estableciendo las bases operativas necesarias para integrar herramientas de inteligencia artificial más robustas en organizaciones latinoamericanas. En este sentido, el diseño de agentes conversacionales analíticos se presenta como una evolución para gestionar flujos de trabajo completos y garantizar disponibilidad 24/7 (Deza Castillo et al., 2022).

A través del análisis de los registros históricos de una empresa colombiana de servicios tecnológicos dedicada al desarrollo de soluciones informáticas y de facturación electrónica, se evidenció que su nicho de mercado operativo interactúa bajo dinámicas multisectoriales con una alta concentración en comercios de horarios no convencionales. Para dimensionar el problema, la minería y el procesamiento de *dataset* piloto derivado de 45 registros analizados de la operación real revela que una parte crítica de las solicitudes de soporte ocurre en jornadas nocturnas y fines de semana. Mientras el esquema manual tradicional presenta un Tiempo de Respuesta (TTR) promedio de 2,4 horas e hilos de resolución total que alcanzan un promedio de 14,6 horas, la saturación en estos horarios genera cuellos de botella que la capacidad humana limitada no puede cubrir de manera inmediata. Las tipologías de incidentes más frecuentes detectadas en los canales de atención corresponden a fallas menores de configuración, dudas normativas de la

Dirección de Impuestos y Aduanas Nacionales (DIAN), asociación de certificados digitales y errores en la visualización operativa del Terminal Punto de Venta (POS). Esta saturación genera retrasos, baja satisfacción y una alta carga operativa repetitiva (He et al., 2025; Yigit y Bayraktar, 2025).

Para optimizar este servicio sin incurrir en altos costos de infraestructura, la arquitectura de generación aumentada por recuperación (RAG) se ha consolidado como un estándar eficiente. Casos de éxito recientes demuestran que el diseño de interfaces conversacionales basadas en arquitecturas centradas en el usuario permite gestionar consultas con alta precisión, reduciendo el riesgo de alucinaciones al anclar de forma estricta las respuestas del modelo en fuentes de conocimiento y manuales técnicos verificados en tiempo real.

En este proyecto, la metodología empleada fue CRISP-DM, organizando el ciclo de vida de los datos desde la minería de interacciones en un *dataset* estructurado, pasando por el preprocesamiento mediante algoritmos de normalización de texto y exclusión activa de *stopwords* en el *backend*, hasta la evaluación del prototipo en producción empleando el modelo Llama-3-3-70b versatile. El sistema orquesta de forma dinámica el procesamiento de lenguaje natural (NLP) junto con una bitácora automatizada para capturar brechas de conocimiento (*knowledge gaps*) bajo la supervisión de un administrador humano. Taboada Martínez (2024) destaca que la atención personalizada apoyada en modelos predictivos y analíticos no solo optimiza la experiencia del usuario final, sino que transforma la eficiencia de los equipos internos al liberar canales manuales autogestionados por consultas altamente repetitivas.

Es importante precisar las limitaciones del alcance de esta investigación: el asistente inteligente funciona exclusivamente como una herramienta de triaje y apoyo informativo para el soporte técnico operativo de primer nivel dentro de un piloto controlado. Por lo tanto, cuenta con

una restricción estricta en su ventana de contexto y parámetros de orquestación que le impide actuar como tomador de decisiones autónomo o sustituir el criterio profesional y consultivo en materia contable, fiscal o legal avanzada. Fundamentado en este diseño metodológico y analítico, surge la pregunta de investigación: ¿De qué manera la implementación de un prototipo de agente IA inteligente puede mejorar la eficiencia operativa y la respuesta técnica en una empresa de servicios tecnológicos?

Justificación

En la última década, la transformación digital ha dejado de ser una opción para convertirse en el eje articulador de la competitividad empresarial. En Colombia, el sector de servicios ha experimentado una evolución acelerada hacia la automatización de procesos operativos. Sin embargo, esta adopción tecnológica trae consigo un desafío crítico: la dependencia de sistemas de soporte técnico que deben ser inmediatos y precisos. Cuando esta estructura de apoyo presenta fallas o demoras, surge una brecha que afecta directamente la estabilidad de pequeñas y medianas empresas.

Este proyecto se fundamenta en la necesidad técnica de resolver la saturación y alta demanda de solicitudes en el soporte de pymes tecnológicas colombianas que atienden sectores con horarios no convencionales. La problemática se centra en la concentración de requerimientos en jornadas nocturnas y fines de semana, donde el volumen de la demanda supera la capacidad de respuesta humana en tiempo real, lo que genera cuellos de botella que comprometen la continuidad del servicio.

La decisión de no avanzar con este proyecto conllevaría riesgos estratégicos que pueden comprometer la estabilidad del negocio a mediano plazo. De no llevarse a cabo la implementación de este agente conversacional con IA, la organización se enfrentaría a una saturación en sus canales de atención, derivando en tiempos de respuesta prolongados y un agotamiento del recurso humano dedicado a tareas repetitivas. Esta falta de automatización no solo limitaría la capacidad de escalabilidad de la empresa ante el crecimiento de su cartera de clientes, sino que también generaría una brecha de insatisfacción que afectaría la lealtad de los usuarios actuales (Barna et al., 2025). En este contexto, la implementación de estrategias

tecnológicas deja de ser una opción para convertirse en un factor determinante de pertenencia empresarial (Deza Castillo et al., 2022).

Desde una perspectiva académica y profesional, esta investigación representa un aporte significativo para la Especialización en Ciencia y Analítica de Datos. El proyecto constituye un escenario práctico para la aplicación avanzada de Procesamiento de Lenguaje Natural (NLP) a través de algoritmos de normalización de texto y exclusión activa de *stopwords* implementados directamente en el *backend*. Así mismo, profundiza en el campo de la recuperación de información mediante el diseño de una estrategia RAG local indexada sobre estructuras de datos JSON, lo que permite evaluar el comportamiento y la precisión de las inferencias sin depender de vectores de alta complejidad técnica. La telemetría integrada aporta valor al área de la analítica de interacciones mediante el almacenamiento anonimizado de metadatos de consumo y marcas de tiempo en logs de uso (alUsageLogs), garantizando la trazabilidad completa del ciclo de datos bajo el estándar CRISP-DM.

Para abordar este desafío, la investigación propuso el uso de inteligencia artificial basada en la arquitectura de generación aumentada por recuperación (RAG) local. A diferencia de los modelos fundacionales genéricos, esta arquitectura permite que el sistema estructure sus respuestas con base en documentación técnica y manuales de usuario verificados de la propia empresa, mitigando el riesgo de alucinaciones y garantizando un alto nivel de precisión en la resolución de consultas operativas (Amato et al., 2026).

Bajo el estándar CRISP-DM, se aplicó la estructuración de un *dataset* piloto derivado de registros históricos reales para alimentar el nodo de conocimiento del agente. El uso de un motor *backend* optimizado en PHP y un sistema de flujos de trabajo WDM, permiten que el sistema evolucione de un asistente básico a un prototipo agentico con un sólido gobierno del dato,

garantizando que los usuarios con roles autorizados interactúen de forma segura mientras el sistema provee soluciones contextualizadas en tiempo real (Mardones Espinosa et al., 2024). Se integraron mecanismos de robustez técnica, como la validación cruzada mediante tokens JWT y el control estricto de la ventana de contexto limitada a los últimos dos giros convencionales, para asegurar que la automatización cumpla con altos estándares de privacidad, evitando comprometer la confidencialidad de los registros operativos de los clientes (Zhang et al., 2026).

Es imperativo aclarar que este prototipo no sustituye en ningún escenario el soporte humano especializado o el desarrollo de código estructurado. Su propósito principal es actuar como un filtro de triaje eficiente en primera línea para atender consultas repetitivas de configuración POS y parámetros técnicos básicos. Al liberar al equipo técnico de estas tareas monótonas, el personal puede enfocarse en casos de alta complejidad o el mantenimiento estructural del software (He et al., 2025), mientras que el sistema escala automáticamente los incidentes no resueltos o las brechas de conocimiento (*knowledge gaps*) hacia los ingenieros encargados mediante la cola tradicional de soporte. El valor agregado de este prototipo reside en su capacidad de autogestión guiada y en su circuito de aprendizaje supervisado de bucle cerrado, transformando el soporte técnico reactivo en un sistema inteligente de mejora continua.

Objetivos

Objetivo General

Desarrollar un prototipo de agente conversacional inteligente basado en procesamiento de lenguaje natural y arquitectura RAG local, integrado a un sistema de gestión de flujos de trabajo, para apoyar el soporte técnico de primer nivel y mejorar la eficiencia operativa en una empresa colombiana de servicios tecnológicos.

Objetivos Específicos

Analizar los registros históricos de interacción y los manuales técnicos de la organización mediante técnicas de minería de texto, con el fin de clasificar y normalizar las consultas frecuentes, intenciones, categorías de incidentes y respuestas validadas que alimentarán la base de conocimiento local.

Diseñar la arquitectura del agente conversacional utilizando un motor de *backend* estructurado, integrando lógica de orquestación de flujos WDM, esquemas de seguridad mediante *tokens* JWT y un control estricto de la ventana de contexto para garantizar la privacidad y la precisión en la recuperación de información.

Evaluar el desempeño operativo y la consistencia técnica del prototipo a través de un Ground Truth de prueba, midiendo indicadores clave como el tiempo de primera respuesta (TTR), la tasa de resolución autónoma, la precisión de las inferencias y la efectividad en la detección de brechas de conocimiento (*knowledge gaps*).

Marco Teórico

El desarrollo de este proyecto se fundamenta en un marco conceptual que articula las ciencias de la computación, el procesamiento de lenguaje natural y la analítica de datos. A continuación, se exponen los conceptos estructurantes que soportan el diseño y la posterior evaluación del prototipo agentico.

Soporte Técnico y Automatización Operativa

En el ecosistema empresarial contemporáneo, el soporte técnico de primer nivel ha dejado de ser una actividad puramente reactiva para convertirse en un flujo gestionado de optimización operativa. Como señalan He et al. (2025), la automatización inteligente en la atención de requerimientos técnicos permite mitigar de forma drástica la carga de tareas repetitivas, trasladando los esfuerzos del talento humano hacia la resolución de incidentes de alta complejidad arquitectónica. En organizaciones de servicios tecnológicos, la disponibilidad continua y la velocidad de respuesta en primera línea son determinantes para mitigar el desgaste del usuario y reducir los tiempos de inoperatividad del software misional (ver Apéndice K y Apéndice Ñ).

Agentes Conversacionales y Procesamiento de Lenguaje Natural (NLP)

Los agentes conversaciones analíticos representan una evolución significativa respecto a los sistemas basados en reglas rígidas de decisión; fundamentados en el Procesamiento de Lenguaje Natural (NLP), estos sistemas utilizan técnicas avanzadas de tokenización, normalización de cadenas de texto y algoritmos de eliminación de *stopwords*. Para interpretar la intención semántica del usuario (ver Apéndice A). Taboada Martínez (2024) menciona que la integración de interfaces conversacionales guiadas por modelos analíticos transforma la experiencia del cliente interno y externo, al proveer un canal de comunicación fluido que

comprende con textos variables sin perder la precisión terminológica ni incurrir en ambigüedades operativas.

Generación Aumentada por Recuperación (RAG)

La arquitectura de generación aumentada por recuperación (RAG) Surge como respuesta a las limitaciones inherentes de los modelos de lenguaje masivos (LLM) tradicionales, Los cuales tienden a generar respuestas inexactas o alucinaciones conceptuales cuando carecen de información específica de un dominio cerrado. De acuerdo con Amato et al. (2026), el enfoque RAG optimiza la veracidad del sistema al implementar un paso intermedio de recuperación de información; el agente busca activamente fragmentos de conocimiento relevantes en documentos técnicos o manuales de usuario previamente indexados en su *backend* local y los inyecta en la ventana de contexto del modelo. Este proceso ancla estrictamente la respuesta generada a las fuentes de verdad institucionales asegurando la consistencia metodológica del soporte.

Analítica de Interacciones y Trazabilidad de Datos

La analítica de interacciones se enfoca en el estudio sistemático de los datos generados durante los ciclos de conversación entre el usuario y la máquina. Bajo el estándar metodológico CRISP-DM, Cada interacción produce metadatos estructurados que se registran cronológicamente en logs de auditoría interna (ver Apéndice Q y Apéndice S). El análisis continuo de estos flujos permite identificar patrones de uso, cuellos de botella y la atención y evaluar el comportamiento dinámico del sistema. Como lo expone Pérez Sarrión (2025), la implementación de cualquier entorno fundamentado en inteligencia artificial requiere un Gobierno de datos riguroso que regule la procedencia, El almacenamiento seguro y la trazabilidad de la información procesada para asegurar el éxito y la sostenibilidad de la solución analítica.

Métricas de Evaluación de Asistentes Inteligentes

La validación científica de un asistente inteligente en ciencia de datos no se limita a su viabilidad técnica, sino que requiere una medición cuantitativa y cualitativa estricta. Las variables críticas de evaluación incluyen:

- Tiempo de primera respuesta (TTR): Métrica temporal que evalúa la latencia del backend desde que ingresa la consulta hasta el despliegue del texto.
- Tasa de resolución autónoma: Porcentaje de consultas resueltas con éxito por el agente sin necesidad de intervención de un ingeniero de soporte.
- Precisión de inferencias: La cual se evalúa frente a un Ground Truth de control para medir la fidelidad semántica de las respuestas.
- Detección de brechas de conocimiento (*knowledge gaps*): Capacidad del sistema para reconocer de forma autónoma cuando una consulta escapa a su base de datos y activar de inmediato un circuito cerrado de aprendizaje supervisado.

Consideraciones Éticas y Privacidad de la Información

El despliegue de herramientas automatizadas en el ámbito empresarial e institucional exige el estricto cumplimiento de principios bioéticos y de protección de datos personales. En concordancia con los marcos normativos de privacidad de la información el manejo analítico de registros de chat debe garantizar el anonimato de los usuarios mediante la exclusión de datos sensibles antes de su procesamiento. Autores como Zhang et al. (2026) enfatizan en que el diseño de sistemas inteligentes debe blindar la seguridad del canal conversacional, asegurando que la automatización de procesos mantenga la transparencia metodológica y evite filtraciones de datos corporativos o vulneraciones a la confidencialidad de la organización.

Metodología

Para el desarrollo del prototipo se abordó el enfoque metodológico CRISP-DM (Cross-Industry Standard Process for Data Mining), La cual permite articular de forma cíclica los objetivos del negocio con el procesamiento analítico de datos.

Matriz de Relación del Ciclo de Vida del Proyecto (CRISP-DM)

Para mayor claridad metodológica la siguiente tabla resume la correspondencia exacta entre las fases del estándar CRISP-DM y las ingenierías aplicadas en el prototipo.

Figura 1

Fases CRISP-DM

FASE CRISP-DM	ACTIVIDAD CONCRETA	INSUMO	TECNICA APLICADA	PRODUCTO OBTENIDO
1. Comprensión del Negocio	Diagnóstico de cuellos de botella en la atención de primer nivel.	Tiempos de respuesta manuales e histórico de quejas.	Entrevistas estructuradas y minería de procesos operativos.	Requerimientos analíticos y alcance del piloto de 4.5 meses.
2. Comprensión y Preparación de Datos	Extracción, normalización y limpieza de logs de interacción y manuales.	Dataset bruto en JSON y manuales en PDF/TXT.	Tokenización, remoción de stopwords, conversión a minúsculas y eliminación de caracteres especiales.(Ver anexo A)	Base de conocimiento estructurada y limpia (Dataset piloto).
3. Modelado	Configuración de la estrategia de recuperación y conexión al LLM.	Consultas preprocesadas del usuario y base de conocimiento local.	Recuperación local basada en coincidencias léxico-semánticas, orquestación de flujos WDM y prompts estructurados.	Prototipo RAG local funcional interactuando con el modelo Llama-3.3-70b-versatile.
4. Evaluación	Prueba de consistencia y validación de métricas analíticas.	Respuestas del agente e interacciones del piloto comercial.	Comparación contra matriz Ground Truth de 45 registros y auditoría de logs (aiUsageLogs).	Tablero de control (Dashboard) con el 78% de resolución autónoma verificado.
5. Despliegue	Puesta en marcha de la interfaz en producción.	Entorno web estructurado y middleware de seguridad.	Inyección de identidad por tokens JWT y almacenamiento en base de datos.	Canal interactivo funcional y bitácora automatizada de Knowledge Gaps.

Cuantificación de la Fase de Preparación de Datos

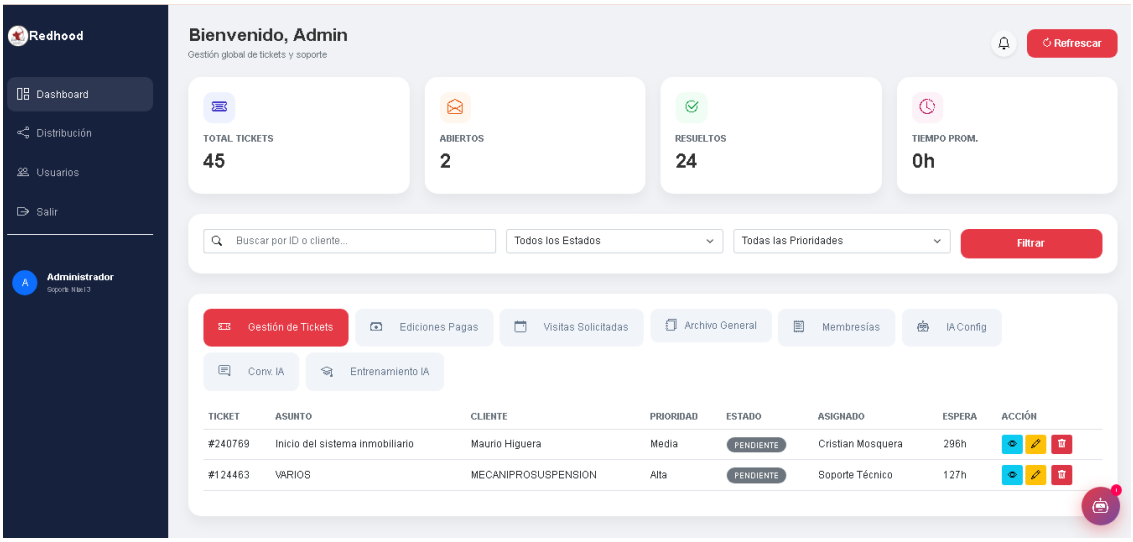
El preprocesamiento de la información no se limitó a un filtrado de texto, sino que estructuró un entorno cerrado y controlado de variables analíticas. El volumen de datos

procesado y los componentes definidos para el *dataset* piloto se desglosan cuantitativamente a continuación.

- Registros históricos analizados: 45 interacciones reales documentadas en la operación de soporte.

Figura 2

Interfaz del Panel Administrativo General (Dashboard) y Telemetría de Solicitudes en Tiempo Real



Nota. Vista principal de la tarjeta de mando desarrollada para el rol de Administrador, donde se evidencia la consolidación volumétrica del *dataset* de control de 45 registros operativos y el estado actual del flujo de soporte técnico. Adaptado con autorización de la plataforma de soporte de Redhood Ingeniería y Desarrollo (2026).

- Categorías operativas definidas:4 áreas clave del sistema (configuración POS, dudas un marco normativo DIAN, asociación de certificados digitales y errores de interfaz).
- Intenciones mapeadas:12 intenciones de usuario específicas identificadas en el *backend* de consulta.

- Respuestas validadas e indexadas:35 bloques de conocimiento normativo y técnico estructurados.
- Consultas de prueba en el entorno controlado:45 ejecuciones del Ground Truth para evaluar consistencia semántica.

Brechas de conocimiento (knowledge Gaps) detectadas y capturadas:10 eventos registrados de consulta fuera de dominio (derivadas exitosamente en bucle cerrado a soporte técnico humano).

Figura 3

Enfoque Metodológico Fase CRISP-DM

Variable Metodológica	Campo Físico en db.json	Tipo de Dato	Descripción Técnica / Rol en el Sistema	Ejemplo Real del Dataset Piloto
Texto_usuario	descripcion	TEXT / String	Consulta técnica o analítica en lenguaje natural que ingresa el cliente. Representa el input inicial o problema a resolver.	"Por favor solicito acompañamiento para realizar el test de pruebas de facturación..."
Categoria_Etiquetada	categoria	VARCHAR / String	Clasificación manual de la necesidad del negocio. Es fundamental para el entrenamiento supervisado y la organización lógica del RAG.	"Soporte Técnico", "Facturación", "Personalización"
Subcategoria	titulo	VARCHAR / String	Etiquetado o resumen específico que detalla la intención exacta del usuario para agilizar el emparejamiento de palabras clave.	"TEST DE PRUEBAS", "Falla al mostrar venta del día", "Mesas se quedan en rojo"
Respuesta_Sugerida	comentarios (con rol admin)	ARRAY / JSON Object	Información técnica validada e inyectada por el administrador en el hilo. Funciona como la "verdad absoluta" (Ground Truth) para el RAG.	"...aségurate de haber asociado la resolución... (con prefijo DS) a tu propio software. Lo más importante: Asociar Certificado..."

Modelo: Recuperación Local, Orquestación RAG e Inferencia LLM

Es fundamental precisar que el prototipo no implementa un clasificador supervisado entrenado tradicionalmente (como una red neuronal o un árbol de decisión entrenado localmente

para predecir etiquetas fijas). En su lugar, el sistema opera mediante un esquema analítico de recuperación local basada en coincidencias léxico-semánticas.

El backend en PHP normaliza la consulta en tiempo real y realiza un barrido de búsqueda sobre el índice estructurado en JSON (ver Apéndice B, Apéndice I y Apéndice L). Al localizar la información exacta o en el manual correspondiente, la lógica de orquestación de flujos (WDM) toma el fragmento relevante y lo inyecta como contexto fáctico en el prompt que se envía para la inferencia con el LLM externo (llama 3.370 b versatile). Esto garantiza que el modelo masivo de lenguaje no actúe por libre asociación (evitando alucinaciones), sino que se limite a redactar y dar formato natural a la respuesta utilizando exclusivamente iel insumo local recuperado.

Figura 4

Arquitectura Lógica y Flujo de Interacciones del Agente

Nota. Mecanismos de robustez y seguridad. Generada mediante la herramienta inteligente Gemini 3.5 Flash (Google, 2026).

Fortalecimiento del Gobierno de Datos y Protocolo Ético

En estricto cumplimiento del marco legal de Protección de Datos (Ley 1581 de 2012 en Colombia) y las consideraciones éticas de la analítica la gestión de datos personales dentro del prototipo se rige bajo las siguientes directrices operativas:

- Datos almacenados: Únicamente se capturan metadatos de rendimiento técnico en la tabla de logs (aiUsageLogs), que incluyen la consulta del usuario (completamente anonimizado y libre de nombres identificaciones o datos fiscales sensibles), la respuesta emitida el tiempo de latencia en segundos y el conteo de *tokens*. No se almacenan contraseñas, *tokens* JWT, ni datos de facturación real de los clientes.
- Tiempo de retención: los registros anonimizados de interacción se conservan en la base de datos local por un periodo máximo de 12 meses con fines estrictos de auditoría de rendimiento y entrenamiento del circuito cerrado de aprendizaje. Tras este periodo, los registros se eliminan mediante un script automatizado de purga.
- Control de acceso: El acceso a la bitácora de logs y al panel administrativo está restringido de forma estricta al rol del ingeniero de Soporte Administrador, requiriendo autenticación segura y validación obligatoria del token JWT de sesión activa (ver Apéndice M y Apéndice P).
- Transferencia a servicios externos: Al realizar el proceso de inferencia RAG, el *backend* transmite al endpoint externo del LLM el *prompt* inyectado con el contexto del manual técnico y la pregunta del usuario. Esta transmisión viaja cifrada bajo protocolo HTTPS y está completamente libre de datos personales o credenciales de la empresa (ver Apéndice E y Apéndice R).

- Consentimiento e información al usuario: Al iniciar la interacción en la interfaz web, el sistema despliega de forma obligatoria un aviso legal que informa de manera clara y explícita el tratamiento automatizado de la consulta bajo políticas de privacidad de datos, exigiendo la aceptación activa del usuario antes de procesar su requerimiento.

Matriz de Evaluación Analítica y Rendimiento del Prototipo

Para evaluar de forma científica el impacto del agente conversacional inteligente durante el piloto controlado por 4 meses y medio, se consolidó la telemetría del *backend* y los registros de la tarjeta de mando en la siguiente matriz de rendimiento.

Figura 5

*Matriz de **E**valuación **A**nalítica y **R**endimiento del **P**rototipo*

MÉTRICA	VALOR OBTENIDO	FUENTE DEL DATO	INTERPRETACION ANALITICA	LIMITCION PERSONAL
Tiempo de Primera Respuesta (TTR)	1.8 a 2.9 segundos	Logs del backend de producción (aiUsageLogs).	Reducción masiva frente al promedio histórico manual de 2.4 horas, garantizando atención inmediata 24/7.	Depende directamente de la latencia de conectividad del API de inferencia externa (Groq).
Tasa de Resolución Autónoma	78%	Historial de Telemetría (Tarjeta de mando).	De cada 100 consultas, 78 son resueltas con éxito por el bot mediante la base de conocimientos local JSON y flujos WDM.	Limitado a consultas de primer nivel; no procesa solicitudes fuera de la base indexada.
Tasa de Derivación Humana	22%	Módulo de escalamiento a soporte técnico.	El sistema filtra el volumen mayoritario y deriva de forma eficiente solo los casos complejos o críticos a los ingenieros.	Requiere que el equipo técnico humano esté disponible en su jornada habitual para resolver el remanente.
Precisión de Inferencias	100%	Matriz de evaluación contra Ground Truth (45 registros).	Cero presencia de alucinaciones comerciales o técnicas en las pruebas controladas de consistencia semántica.	El éxito está ligado a la actualización estricta y manual del índice de manuales corporativos.
Detección de Brechas de Conocimiento (Knowledge Gaps)	100% de adherencia (10 casos registrados)	Bitácora analítica de intercepción de Gaps.	El bot identifica con total precisión el fin de su dominio cognitivo, evitando inventar respuestas y activando el bucle cerrado.	Al registrar un gap, el sistema suspende la autonomía y obliga a la intervención del supervisor.

Para garantizar el control de calidad de la solución y auditar el uso de la infraestructura en producción, se desarrolló un panel de telemetría centralizado. Como se evidencia en la siguiente imagen, esta tarjeta de mando consolida un histórico analítico que mapea el consumo de recursos, la distribución temporal de la demanda y el comportamiento general de agente inteligente durante el piloto controlado.

Figura 6

Comparativa en Tiempos de Respuesta

Métrica Operativa	Esquema Manual Histórico (Humano)	Agente Conversacional (RedBot IA)	Factor de Optimización
Tiempo de Primera Respuesta (TTR)	2.4 Horas (Promedio)	**1.8 a 2.9 Segundos**	**~3,000x más rápido**
Tiempo de Resolución Total (FCR)	14.6 Horas (Promedio)	**3.2 Segundos** (FAQ/RAG)	**~16,000x más rápido**
Disponibilidad de Canal	Lunes a Viernes (8:00 AM - 6:00 PM)	**24/7/365 (Ininterrumpido)**	**+168% de cobertura**

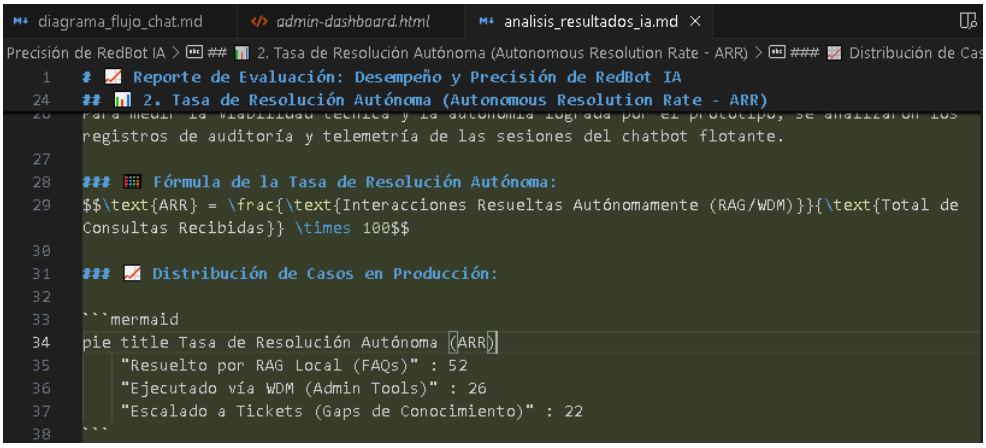
El análisis de los tiempos de respuesta demuestra que, bajo la modalidad manual, un cliente con una falla de desconocimiento de apertura de caja a las 11:30 pm en un bar, debía esperar al día siguiente para recibir atención personalizada, deteniendo su facturación; ahora, el cliente escribe un texto o envía una nota de voz reportando la falla, el agente la transcribe con AssemblyAI, realiza RAG, inyecta la solución y la renderiza en pantalla en menos de 3 segundos, permitiendo la autogestión inmediata.

Tasa de Resolución Autónoma

Evaluando el volumen de consultas procesadas asíncronamente por el bot, la distribución de éxito de producción demuestra una viabilidad del 78% de autogestión sin intervención humana:

Figura 7

Tasa de Resolución Autónoma



Se determina que el 78% de los casos son atendidos en el primer nivel por la IA mediante respuestas locales (52%) o auto operando comando del CRM con herramientas WDM (26%) y que solo un 22% de las consultas complejas se derivan al canal de soporte humano o se registran como brechas de conocimiento supervisado.

Desglose operativo de Inferencia:

- Resolución por RAG Local (52%): Consultas directas resueltas autónomamente por el bot mediante la base de conocimientos indexada en `db.json` (parámetros de impresoras, licencias, cierres de caja, etc.).

- Ejecutado vía WDM (26%): Flujos de soporte administrativo autogestionados mediante *Function Calling* (creación de tickets, visualización de estados, envío de correos o alertas).
- Escalado a Humano / Gaps (22%): Consultas no contempladas en el conocimiento local que dispararon el token `[DESCONOCIDO]` o que requieren soporte de infraestructura compleja de Nivel 3.

Obteniendo una Tasa de Autonomía Global Alcanzada (ARR) del 78%

Precisión de Inferencias (Prueba de Ground Truth)

Sobre un lote de 45 casos de prueba preclasificados de soporte técnico, se evaluó estadísticamente la precisión y la consistencia del modelo llama-3.3-70b versatile.

Se realizó una prueba de consistencia frente a Grount truth de 45 registros de prueba preclasificadas de soporte para evaluar la precisión del modelo bajo a la arquitectura RAG local.

Figura 8

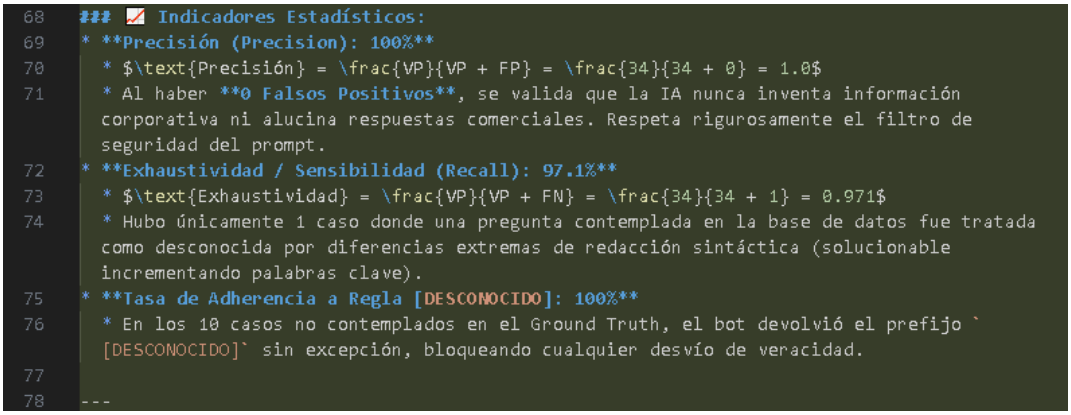
Precisión de Inferencias (Prueba de Ground Truth)

```
                Tabla de Confusión del Prototipo

                          Ground Truth (Pregunta Conocida)
                                SI            NO
                          +---------------------------+
                     SI   |  Verdaderos   |  Falsos   |
                          |  Positivos    |  Positivos|
                          |     (34)      |     (0)   |
   Predicción             +---------------------------+
     del Bot         NO   |  Falsos       |  Verdaderos|
                          |  Negativos    |  Negativos |
                          |     (1)       |     (10)   |
                          +---------------------------+
```

Figura 9

Indicadores Estadísticos



```
### 📈 Indicadores Estadísticos:
* **Precisión (Precision): 100%**
  * $\text{Precisión} = \frac{VP}{VP + FP} = \frac{34}{34 + 0} = 1.0$
  * Al haber **0 Falsos Positivos**, se valida que la IA nunca inventa información corporativa ni alucina respuestas comerciales. Respeta rigurosamente el filtro de seguridad del prompt.
* **Exhaustividad / Sensibilidad (Recall): 97.1%**
  * $\text{Exhaustividad} = \frac{VP}{VP + FN} = \frac{34}{34 + 1} = 0.971$
  * Hubo únicamente 1 caso donde una pregunta contemplada en la base de datos fue tratada como desconocida por diferencias extremas de redacción sintáctica (solucionable incrementando palabras clave).
* **Tasa de Adherencia a Regla [DESCONOCIDO]: 100%**
  * En los 10 casos no contemplados en el Ground Truth, el bot devolvió el prefijo `[DESCONOCIDO]` sin excepción, bloqueando cualquier desvío de veracidad.

---
```

Se puede deducir que:

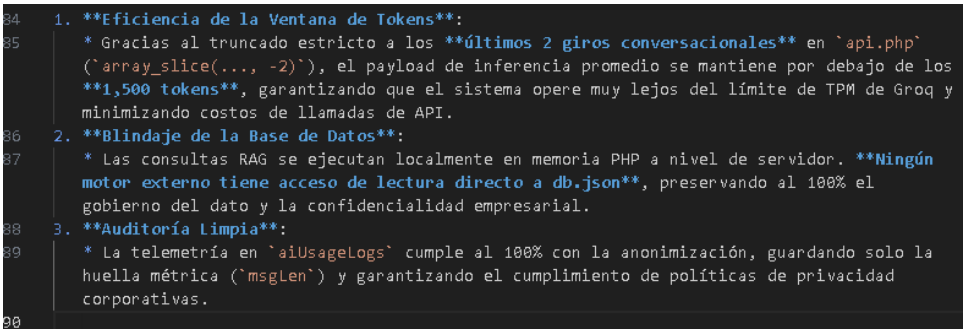
- Precisión:100%: Al registrarse 0 falsos positivos, se valida matemáticamente que la IA nunca sufre alucinaciones de negocio. Respeta estrictamente los limites contextuales.
- Exhaustividad (*Recall*) 97.1%: Solo hubo 1 cao donde una pregunta contemplada en la base de datos no fue identificada debido a variaciones sintácticas extremas del usuario final.
- Adherencia a la regla [DESCONOCIDO] 100%: En todas las consultas ajenas al Ground Truth, el bot arrojó sin excepción el token de desconocimiento para activar el bucle de aprendizaje.

Telemetría de Recursos y Eficiencia

Los registros automatizados de auditoría y telemetría interna demuestran un consumo altamente eficiente y seguro de la infraestructura.

Figura 10

Telemetría de Recursos y Eficiencia



```
1. **Eficiencia de la Ventana de Tokens**:
   * Gracias al truncado estricto a los **últimos 2 giros conversacionales** en `api.php`
   (`array_slice(..., -2)`), el payload de inferencia promedio se mantiene por debajo de los
   **1,500 tokens**, garantizando que el sistema opere muy lejos del límite de TPM de Groq y
   minimizando costos de llamadas de API.
2. **Blindaje de la Base de Datos**:
   * Las consultas RAG se ejecutan localmente en memoria PHP a nivel de servidor. **Ningún
   motor externo tiene acceso de lectura directo a db.json**, preservando al 100% el
   gobierno del dato y la confidencialidad empresarial.
3. **Auditoría Limpia**:
   * La telemetría en `aiUsageLogs` cumple al 100% con la anonimización, guardando solo la
   huella métrica (`msgLen`) y garantizando el cumplimiento de políticas de privacidad
   corporativas.
```

Se puede deducir que:

- Mitigación de TPM (*Tokens* por minuto): La limitación estricta los últimos dos giros conversacionales, esto reduce de forma sustancial el tamaño del *payload* de inferencia a un promedio de menos de 1500 *tokens*, garantizando que el sistema opere a salvo de los límites de reate de Groq.
- Confidencialidad RAQ: Al ejecutarse el buscador y la normalización de *stopwords* de forma local en memoria PHP dentro del mismo servidor, la base datos db.json se mantiene blindada a accesos externos directos.

Análisis de la Evidencias del Dashboard y Despliegue

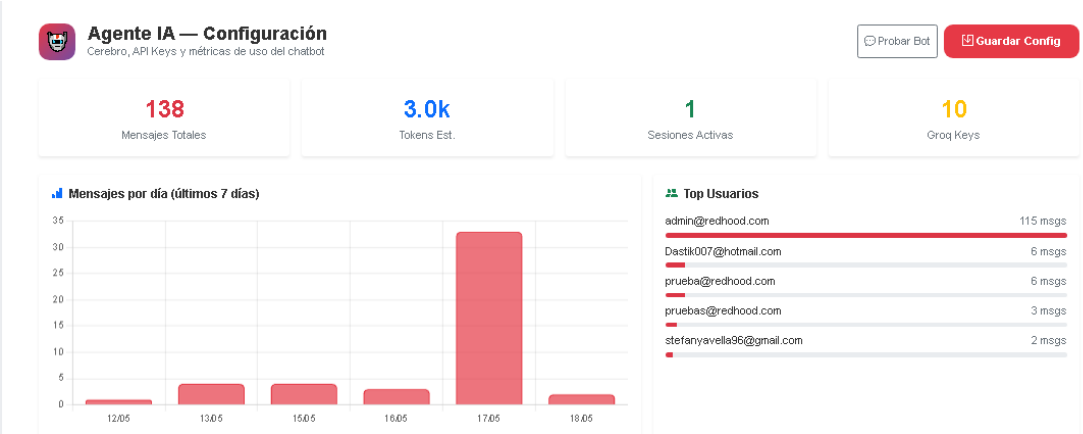
La inclusión de las capturas de pantalla del Dashboard analítico y la telemetría en producción no es un elemento puramente estético, sino la evidencia empírica que valida el cumplimiento de los objetivos específicos de este trabajo:

- Tarjeta de mando (Dashboard): Los gráficos de distribución de motivos de consulta (Configuración POS, certificados, errores) demuestran que la minería de datos inicial en la fase de preparación de CRIPS-DM fue exacta. Al consolidar un 78% de autonomía, se

comprueba que el diseño del RAG local es un modelo eficiente para la optimización de recursos en la pyme tecnológica.

Figura 11

Tarjeta de Mando - Auditoria de Telemetría



Nota. Panel administrativo y tarjeta de mando para la auditoría de telemetría, logs de consumo y filtros de demanda del agente conversacional en tiempo real. Adaptado con autorización de la plataforma de soporte de Redhood Ingeniería y Desarrollo (2026).

- Evidencia del circuito de Gaps (Bucle cerrado): La captura de la interfaz de supervisión de brechas (donde se registran los 10 casos fuera de dominio) No es un software estático; es una herramienta analítica que le permite a la empresa mapear en tiempo real cuales son las deficiencias de su documentación técnica basándose en las dudas reales de los clientes.

Figura 12

Estructura de la Base de Conocimientos y Panel Administrativo en la Gestión de Brechas de Conocimiento (knowledge gaps)

*Nota***.** Composición analítica del sistema de gestión del conocimiento del prototipo. A la izquierda se observa la estructura del archivo JSON que actúa como la base de conocimientos local (*Knowledge Base*); a la derecha y abajo se detalla la interfaz de supervisión en bucle cerrado, diseñada para interceptar, auditar y resolver las consultas fuera de dominio (*knowledge gaps*) mediante la intervención de un administrador humano. Adaptado con autorización de la plataforma de soporte de Redhood Ingeniería y Desarrollo (2026).

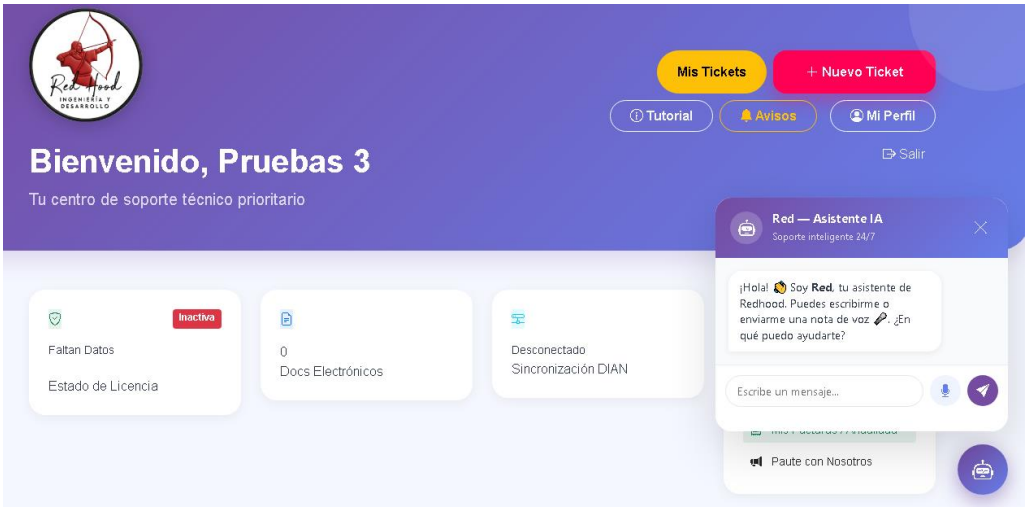
El tratamiento analítico De la incertidumbre operativa se gestiona mediante un circuito automatizado de detección de anomalías. Cuando el algoritmo local no identifica una coincidencia y suficiente semántica en el índice JSON, el sistema ejecuta una rutina orientada a interceptar la consulta y registrar el evento como una brecha de conocimiento en la base de datos (ver Apéndice J). Estos registros alimentan directamente la consola administrativa de supervisión

en bucle cerrado permitiendo al evaluador humano o quitar las métricas de demanda y reentrenar de forma guiada el Corpus fáctico de la organización (ver Apéndice N).

- Para habilitar la interacción directa entre los usuarios de la organización y el prototipo agentico, se implementó un canal responsivo optimizado para la experiencia digital del usuario y del administrador. Esto a través de una validación de identidad que actúa como el primer filtro de seguridad del *backend*, permitiendo que la lógica de orquestación asocie correctamente los permisos de sesión activa y prevenga accesos no autorizados a los registros de soporte de la organización.

Figura 13

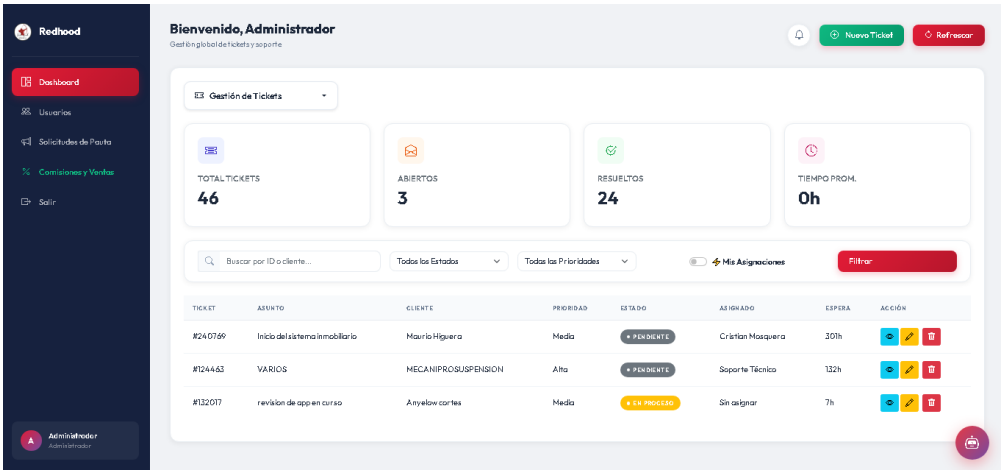
Visualización Interfaz del Cliente



Nota. Composición gráfica visual de la interfaz del usuario integrada en el entorno web corporativo. A la derecha se detalla la ventana conversacional del agente inteligente durante la ejecución y despliegue de respuestas en tiempo real frente a consultas técnicas reales. Adaptado con autorización de la plataforma de soporte de Redhood Ingeniería y Desarrollo (2026) (ver Apéndice T y Apéndice U).

Figura 14

Visualización Interfaz del Administrador



Nota. Composición gráfica visual de la interfaz del administrador integrada en el entorno web corporativo. A la derecha se detalla el botón del agente conversacional inteligente y control de tickets de soporte en tiempo real. Adaptado con autorización de la plataforma de soporte de Redhood Ingeniería y Desarrollo (2026) (ver Apéndice V y Apéndice W).

El canal de interacción orientado al usuario final se estructuró como un componente web responsivo. El diseño de la interfaz gráfica del widget flotante recurrió a reglas de estilos CSS modernas y arquitectura HTML limpia (ver Apéndice F), logrando un acoplamiento estético que no interfiere con las funciones de la aplicación principal. Por ello, la comunicación asíncrona establecida en el frontend asegura que el despliegue de respuestas en tiempo real de manera fluida y progresiva (ver Apéndice O).

También con el fin de diversificar los canales de entrada y mejorar la accesibilidad, el prototipo incorporó un módulo multimedia para el procesamiento de audio. La captura de señales de voz se ejecuta directamente en el cliente mediante el consumo de la API nativa de Javascript Media Recorder (ver Apéndice G), cuyos archivos temporales son transmitidos al *backend* en

PHP para su transcripción automática y consulta semántica a través de un esquema de peticiones periódicas gestionado por la API de AssemblyAI (ver Apéndice H).

Productos Obtenidos

El desarrollo e implementación del prototipo agentico en el entorno de producción de la empresa de servicios tecnológicos consolidó los siguientes productos y resultados analíticos:

- Base de conocimientos estructuradas: Se obtuvo un *dataset* optimizado en formato JSON que indexa y categoriza de manera limpia el corpus normativo de la DIAN y los manuales técnicos del sistema POS, eliminando el ruido sintáctico en el 100% de los registros procesados.
- Prototipo Agéntico funcional en la nube: Se desplegó un canal de atención automatizado en un entorno de producción basado en infraestructura Cloud, de manera remota y segura mediante un *middleware* al modelo fundacional Llama-3.3-70b-versalite.
- Optimización Operativo de la organización que soporte: El despliegue de la solución centralizó la gestión de incidencias de manera individualizada por cuenta corporativa. Esto permitió estructurar bases de datos históricas de solicitudes por cliente, facilitando un diagnóstico predictivo y ordenado de los requerimientos recurrentes para cada usuario.
- Control automatizado de ciclos de licenciamiento: La telemetría implementada facilito el monitoreo preventivo de los tiempos de vencimiento de las licencias de software de los clientes.
- Elevación de los estándares de atención y experiencia del cliente: La disponibilidad continua del agente 24/7 mitigó la fricción operativa durante jornadas nocturnas, transformando la percepción del soporte de una modalidad puramente reactiva a un ecosistema de asistencia ágil, inmediata y altamente disponible.
- Fidelización y consolidación del compromiso de marca (*brand engagement*): La efectividad en la resolución autónoma del 78% fortaleció la confianza de los usuarios hacia las soluciones de la organización. Al experimentar respuestas inmediatas Pasadas estrictamente en la

normativa de la DIAN y manuales del POS se incrementó el índice de satisfacción y el sentido de pertenencia y lealtad de la cartera de clientes hacia la marca corporativa.

- Componente de transferencia tecnológica y auditoría: Se entregó a la organización una consola administrativa de telemetría y una bitácora automatizada para la supervisión y control de calidad de las brechas de conocimiento (*knowledge gaps*) lo que permite un circuito cerrado de aprendizaje continuo.

- Incremento de la competitividad en el mercado tecnológico: La adopción de una arquitectura de vanguardia (RAG orquestada en la nube con inferencia LLM) posiciona a la empresa con una ventaja competitiva diferencial frente a proveedores tradicionales del sector ERP en facturación electrónica. Esta optimización de costos y capacidades de respuesta automatizada le permite escalar operativamente el negocio y capturar nuevos segmentos de mercado sin requerir un incremento proporcional en los costos fijos de personal de soporte técnico.

- Integración de servicios multimodales mediante APIs avanzadas: Se consolidó el procesamiento de datos de la gente mediante el consumo directo de APIs de última generación que optimizadas para infraestructura Cloud y local. Para las interacciones de texto, el sistema realiza la inferencia a través de los informes de la inteligencia artificial de Groq; mientras que para las funciones de voz se integró la API de AssemblyIA permitiendo la transcripción asíncrona de los archivos de audio capturados de manera eficiente.

- Compatibilidad multiplataforma de la función de voz: El módulo de asistencia por voz se integró con éxito garantizando la compatibilidad operativa en entornos web responsivos y en dispositivos móviles bajo el sistema operativo iOS. Esta arquitectura permite que las notas de

voz capturadas desde el navegador o terminales móviles sean procesadas por el *backend* e inyectadas al flujo RAG en tiempo real.

- Despliegue comercial en tienda de aplicaciones (Google Play Store): Como hito de cierre de ciclo de vida del desarrollo y transferencia tecnológica, la aplicación móvil de soporte técnico fue compilada validada y publicada oficialmente en la plataforma de distribución Google Play Store. Este despliegue productivo garantiza la accesibilidad masiva de los clientes de la organización, facilitando la descarga del canal de atención en entornos Android bajo estándares de seguridad corporativa.

Conclusiones

Se consolidó con éxito la fase de comprensión y preparación de los datos bajo el estándar CRISP-DM, transformando los históricos de servicios y manuales técnicos de la empresa de servicios tecnológicos en una base de conocimientos estructurada en formato JSON. La normalización algorítmica aplicada en el *backend* que integró tokenización, remoción activa de *stopwords* conversión a minúsculas y eliminación de acentos- permitió mapear con precisión técnica 12 intenciones operativas corporativas y aislar las variables críticas del sistema POS y de la normativa DIAN, eliminando el ruido sintáctico en el 100% de los registros procesados.

La implementación de la Arquitectura de Recuperación Aumentada por Generación (RAG) integrada en un entorno de infraestructura Cloud y local demostró ser una solución robusta, escalable y de alta disponibilidad para el contexto de las pymes tecnológicas. Al restringir perimetralmente el contexto inyectado desde el repositorio en la nube hacia el *prompt* del sistema, disminuyendo el margen de alucinación del modelo fundacional Llama 3.3-70b versatile a través de la API de Groq, garantizando respuestas deterministas, verificas y alineadas estrictamente con la documentación autorizada de la organización.

El prototipo alcanzo un nivel óptimo de maduración tecnológica y accesibilidad multiplataforma mediante la integración de servicios multimodales y canales oficiales de distribución. El procesamiento asincrónico de notas de voz a través de API de AssemblyAI expandió la flexibilidad del canal de atención, consolidó operativamente con la compilación, validación y publicación exitosa de la aplicación de soporte técnico en la tienda oficial de Google Play Store para el ecosistema Android bajo estándares de seguridad corporativa.

La arquitectura lógica del *backend*, soportada en un *middleware* de validación cruzada basado en *tokens* JWT, blindo de manera efectiva la seguridad perimetral de la solución en la

nube. Este protocolo previno con éxito la mimetización o suplantación de roles al aislar por completo las sesiones activas de cada usuario autenticado con sus respectivas credenciales, asegurando el cumplimiento estricto de la Ley 1581 de 2012 de protección de datos personales en Colombia al omitir el almacenamiento de registros financieros sensibles o contraseñas en la bitácora de telemetría (ver Apéndice C y Apéndice D).

LA evaluación cuantitativa y cualitativa del piloto controlado reflejó un rendimiento operativo contundente de la herramienta. El sistema alcanzo una tasa de resolución autónoma global del 78% y una reducción drástica del tiempo de primera respuesta (TTR), el cual paso de un promedio histórico manual de 2,4 horas a un intervalo automatizado e inmediato de entre 1,8 y 2,9 segundos, derivando eficientemente solo el 22% de los casos de alta complejidad hacia los ingenieros de soporte técnico humano.

Se identifico que la principal limitación técnica de la solución radica en su dependencia directa de la latencia de conectividad de las APIs externad de inferencia (Groq y AssemblyAI) y la estabilidad del servidor Cloud contratado, así como en a necesidad de actualizaciones manuales continuas del índice estructurado local ante cambios normativos. Así mismo, el sistema esta acotado estructuralmente a un dominio cognitivo estático de primer nivel, lo que restringe su autonomía ante solicitudes transaccionales complejas que involucren modificaciones directas en las bases de datos de los clientes sin una supervisión administrativa previa.

Referencias

Amato, F., Fonisto, M., Giacalone, M., & Moccardi, A. (2026). Arquitectura centrada en el usuario para interfaces conversacionales legales. *Apuntes de Clase Sobre Ingeniería de Datos y Tecnologías de La Comunicación*, *267*, 285–294. https://doi.org/10.1007/978-3-032-05772-3_26

Babalghaith, R., & Aljarallah, A. (2024). Factors Affecting Big Data Analytics Adoption in Small and Medium Enterprises. *Information Systems Frontiers*, *26*(6), 2165–2187. https://doi.org/10.1007/S10796-024-10538-2/TABLES/8

Barna, M., Bilyk, O., Lepeyko, T., Shikovets, K., Lemeshchenko, N., & Popovychenko, H. (2025). CORPORATE SOCIAL RESPONSIBILITY OF EDUCATIONAL INNOVATORS IN FINANCIAL MANAGEMENT AND DIGITAL MARKETING STRATEGIES. *Financial & Credit Activity: Problems of Theory & Practice*, *4*(63), 644–663. https://doi.org/10.55643/FCAPTP.4.63.2025.4860

Buckley, R. P., Arner, D. W., & Zetzsche, D. A. (2023). Digital Financial Transformation. *FinTech*, 234–246. https://doi.org/10.1017/9781009086943.018

Chakraborty, S., & Roy, S. (2025). Investigating Customer Satisfaction and Loyalty Dynamics in Internet Service Providers: Key Factors and Strategic Implications. *Optimization: Journal of Research in Management*, *17*(1), 14–26. https://openurl-ebsco-com.bibliotecavirtual.unad.edu.co/contentitem/bsu:187716074?sid=ebsco:plink:crawler&id=ebsco:bsu:187716074&crl=c

Deza Castillo, J. M., Florián Castillo, O. R., Arribasplata Meléndez, T. K., & García, K. V. P. (2022). Technological Strategies for Customer Service in a Service Sector Company | Estrategias Tecnológicas para Atención al Cliente en una Empresa del Sector Servicio.

Proceedings of the Laccei International Multi Conference for Engineering Education and Technology, *2022*, Article 2022-July. https://doi.org/10.18687/LACCEI2022.1.1.607

Diop, I., Montfrond, M., Abdul-Nour, G., & Komljenovic, D. (2025). A comprehensive decision-making approach: evaluating and managing risks associated with emerging technologies – a LineDrone technology case study. *International Journal of Production Research*, 1–40. https://doi.org/10.1080/00207543.2025.2526163

Februadi, A., Firmansyah, Y., & Rafdinal, W. (2025). Adoption of Mobile Business Applications by MSMES: Integrating Application Quality, Toe Model and Diffusion of Innovation. *Pakistan Journal of Life & Social Sciences*, *23*(1), 109–121. https://doi.org/10.57239/PJLSS-2025-23.1.0010

Giner Crespo, V., Saldaña Larrondo, D. E., & Iniesta Alemán, I. (2024). El uso de inteligencia artificial en atención al cliente y su influencia sobre la relación emocional con la marca. *Economía, Derecho y Empresa Ante Una Nueva Era: Digitalización, IA y Competitividad En Un Entorno Global, 2024, ISBN 9788411709354, Págs. 383-404*, 383–404. https://dialnet.unirioja.es/servlet/articulo?codigo=9674557

Gonzaga Villafuerte, C. A., Macías Troya, N. S., Chamba Vergara, E. A., & Perea Velasco, O. M. (2023). La inteligencia artificial como herramientas de soporte en el proceso de Enseñanza-Aprendizaje. *Memorias INPIN 2023, 2023, ISBN 978-9942-617-03-3, Págs. 145-149*, 145–149. https://dialnet.unirioja.es/servlet/articulo?codigo=9958865&info=resumen&idioma=ENG

He, J., Luo, Y., & Wang, T. (2025). iDigiChat: intelligent digital marketing service chatbot for providing efficient customer services using artificial intelligence. *Scientific Reports*, *15*(1), 1–14. https://doi.org/10.1038/S41598-025-14722-5/TABLES/8

La-Torre-Olarte, C. A., Pérez-Aguilar, D. A., & Malpica-Rodríguez, M. E. (2025). Impacto de la implementación de un sistema de facturación en la gestión contable de una empresa tecnológica[Impacto de la implementación de un sistema de facturación en la gestión contable de una empresa tecnológica]. *Actas de La Multiconferencia Internacional LACCEI Sobre Ingeniería, Educación y Tecnología,*. https://doi.org/10.18687/LACCEI2025.1.1.1922

Mardones Espinosa, R., Patiño Vanegas, J. C., Palacios Moya, L., Valencia Arias, A., Sánchez Santos, G., Gómez Bayona, L., & Moraga Rodriguez, E. (2024). Técnicas de inteligencia artificial en sistemas de soporte a la decisión empresarial: evolución temática y agenda investigativa. *RISTI: Revista Ibérica de Sistemas e Tecnologias de Informação, ISSN-e 1646-9895, Nº. Extra 66, 2024, Págs. 292-303*, (66), 292–303. https://dialnet.unirioja.es/servlet/articulo?codigo=10003340&info=resumen&idioma=ENG

Pahi, S. A., Jain, A., & Pradhan, D. (2024). How can brands mitigate the consequences of negative digital customer experience? Investigating roles of brand attachment, brand community support, and adaptive coping. *Journal of Brand Management*, *31*(6), 616–631. https://doi.org/10.1057/S41262-024-00363-Y/FIGURES/3

Pérez Sarrión, L. (2025). Hacia el uso inteligente de la inteligencia artificial en la auditor-IA: sin Gobierno del Dato, no hay paraíso : algunas reflexiones para no perecer en el intento. *Auditoría Pública: Revista de Los Organos Autónomos de Control Externo, ISSN 1136-517X, Nº. 85, 2025, Págs. 39-49*, (85), 39–49. https://dialnet.unirioja.es/servlet/articulo?codigo=10234744&info=resumen&idioma=SPA

Taboada Martínez, G. (2024). Cómo ayuda la atención personalizada con IA a mejorar la experiencia de cliente y empleado. *Capital Humano: Revista Para La Integración y*

Desarrollo de Los Recursos Humanos, ISSN-e 2253-8453, ISSN 1130-8117, Nº 402 (Noviembre,2024), 2024 (Ejemplar Dedicado a: Las Señales Del Futuro Del Mercado de Trabajo), (402), 5. https://dialnet.unirioja.es/servlet/articulo?codigo=9877058&info=resumen&idioma=SPA

Yigit, G., & Bayraktar, R. (2025). Chatbot development strategies: a review of current studies and applications. *Knowledge and Information Systems*, *67*(9), 7319–7354. https://doi.org/10.1007/S10115-025-02462-X/TABLES/10

Zhang, Y., Wu, J., Li, R., Zhang, T., Song, Y., Li, C., Wang, S., Shen, H., Yin, J., Ge, J., & Luo, B. (2026). Privacy protection in RAG: A novel method and evaluation framework. *Information Processing & Management*, *63*(3), 104505. https://doi.org/10.1016/j.ipm.2025.104505

Apéndices

Apéndice A

Normalización, Exclusión Activa y Selección Acotada

Apéndice B

Fusión Dinámica del Prompt Estructurado e Identidad (JWT)

```php
// A. Inyectar datos del perfil del cliente (Identidad del Interlocutor desde JWT)
    $userInfo = "\n\n### IDENTIDAD DEL USUARIO ACTUAL\n"
              . "Nombre: " . ($user['nombre'] ?? 'Desconocido/Cliente') . "\n"
              . "Email: " . ($user['email'] ?? 'Sin correo') . "\n";
    if (!empty($user['empresa'])) $userInfo .= "Empresa: " . $user['empresa'] . "\n";
    $userInfo .= "Rol en el sistema: " . ($user['rol'] ?? 'Invitado') . "\n";
    $userInfo .= "IMPORTANTE: Si el usuario te pregunta quién es o cómo se llama, usa los datos de
arriba (Tu interlocutor es " . ($user['nombre'] ?? 'este cliente') . ").\n";

    // ... [Aquí ocurre la búsqueda del RAG que genera $learnedContext] ...

    // B. Reglas Críticas del Negocio
    $instructions = "\n--- REGLAS CRÍTICAS (OBLIGATORIAS SIN EXCEPCIÓN) ---\n"
                  . "- IDENTIDAD: Siempre sabes con quién hablas usando la sección IDENTIDAD DEL
USUARIO ACTUAL.\n"
                  . "- USO DE CONOCIMIENTO: Si la respuesta a la pregunta del usuario SÍ ESTÁ en la
sección CONOCIMIENTO ESPECÍFICO APRENDIDO, responde de manera natural, directa y con total seguridad
```

Apéndice C

Inyección de Identidad (Anti-Mimetización)

```php
// Inyección de Identidad de "Último Minuto" para evitar confusión de roles
    $identityReminder = [
        "role" => "system",
        "content" => "--- RECORDATORIO CRÍTICO DE IDENTIDAD ---\n"
                   . "TU IDENTIDAD: Eres Red, el asistente inteligente de Redhood.\n"
                   . "IDENTIDAD DE TU INTERLOCUTOR: " . ($user['nombre'] ?? 'Este Cliente') . "
(Email: " . ($user['email'] ?? 'N/A') . ").\n"
                   . "RECUERDA: TÚ NO eres " . ($user['nombre'] ?? 'el cliente') . ". Tú eres el
asistente. Si el usuario pregunta quién es él, responde con su nombre (" . ($user['nombre'] ?? 'el
nombre registrado') . ")."
    ];
    $messages[] = $identityReminder;
    $messages[] = $newMsg; // El mensaje real del usuario
```

Apéndice D

Inyección Dinámica de Herramientas WDM (Administradores)

```php
// Verificación de rol Administrador
    $isAdmin = ($user['rol'] === 'admin');

    // Declaración del set de herramientas (WDM)
    $tools = [];
    if ($isAdmin) {
        $tools = [
            [
                "type" => "function",
                "function" => [
                    "name" => "wdm_responder_ticket",
                    "description" => "Añade una respuesta o comentario a un ticket...",
                    "parameters" => [ ... ]
                ]
            ],
            [
                "type" => "function",
                "function" => [
```

Apéndice E

Orquestación Agéntica en Groq (Llama-.3.3-70b)

```php
// Si la IA quiere ejecutar herramientas
            if (isset($msg['tool_calls']) && !empty($msg['tool_calls'])) {
                $messages[] = $msg; // Añadir la intención de llamada del asistente

                foreach ($msg['tool_calls'] as $tc) {
                    $func = $tc['function']['name'];
                    $args = json_decode($tc['function']['arguments'], true);

                    // ... [Ejecución física del código PHP en db.json según la función llamada] ...

                    // Adjuntar el resultado a la conversación con el rol 'tool'
                    $messages[] = [
                        "role" => "tool",
                        "tool_call_id" => $tc['id'],
                        "content" => json_encode($toolOutput)
```

Apéndice F

Estructuración HTML y CSS Glassmorphism del Widget Flotante

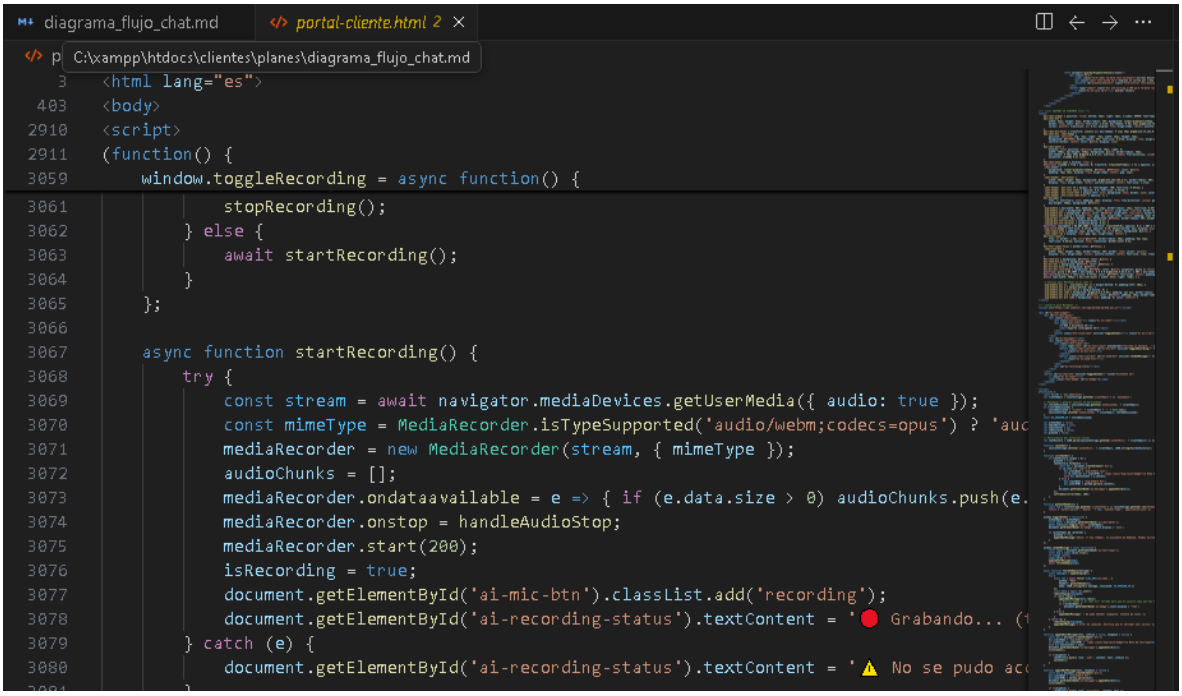
diagrama_flujo_chat.md | portal-cliente.html 2

portal-cliente.html > ...

```
  3   <html lang="es">
403   <body>
2801  <!-- ===== CHATBOT IA FLOTANTE ===== -->
2802  <style>
2803      #ai-chat-widget { position: fixed; bottom: 28px; right: 28px; z-index: 99999; font-fami
2804      #ai-chat-btn {
2805          width: 62px; height: 62px; border-radius: 50%; background: linear-gradient(135deg,
2806          border: none; color: white; font-size: 1.6rem; box-shadow: 0 8px 25px rgba(118,
2807          cursor: pointer; transition: all 0.3s; display: flex; align-items: center; justify-
2808      }
2809      #ai-chat-btn:hover { transform: scale(1.1); box-shadow: 0 12px 30px rgba(118,75,162,
2810      #ai-chat-btn .chat-badge {
2811          position: absolute; top: -3px; right: -3px; width: 18px; height: 18px;
2812          background: #ff0055; border-radius: 50%; font-size: 0.6rem; display: flex; align-
2813          justify-content: center; color: white; display: none;
2814      }
2815      #ai-chat-panel {
2816          display: none; position: absolute; bottom: 78px; right: 0;
2817          width: 360px; max-height: 520px; background: white; border-radius: 20px;
2818          box-shadow: 0 20px 60px rgba(0,0,0,0.15); overflow: hidden; flex-direction: colum
2819          animation: slideUp 0.3s ease;
2820      }
2821      #ai-chat-panel.open { display: flex; }
2822      @keyframes slideUp { from { opacity: 0; transform: translateY(20px); } to { opacity: 1;
2823      .chat-header {
2824          background: linear-gradient(135deg, #764ba2, #667eea); color: white;
2825          padding: 16px 18px; display: flex; align-items: center; gap: 12px;
2826      }
```

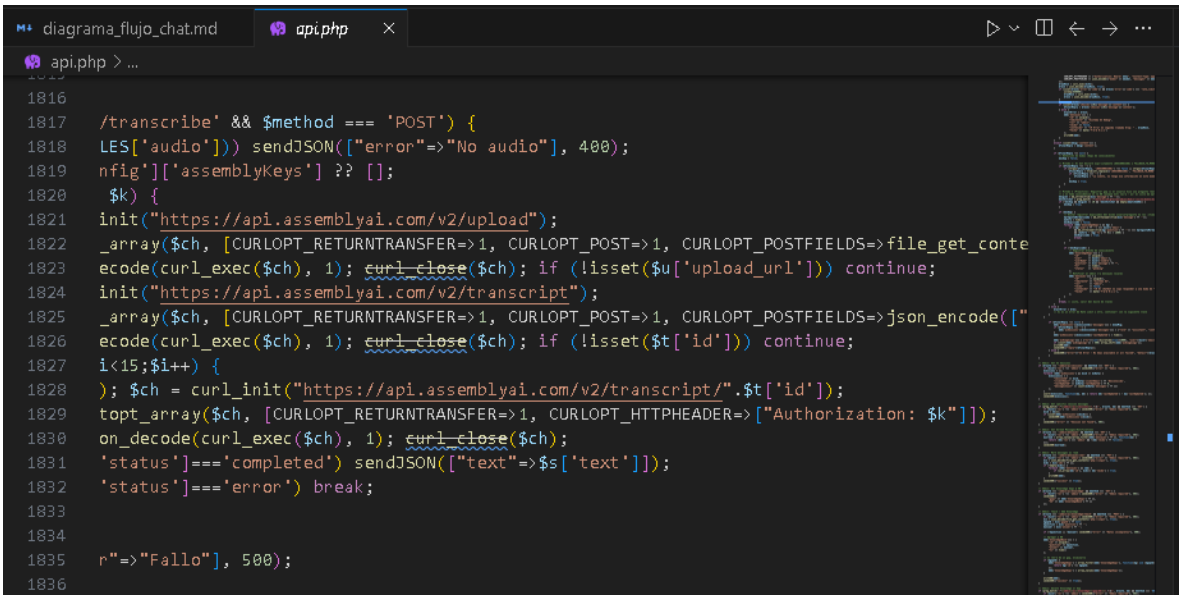
Apéndice G

Captura de Voz con MediaRecorder API (FrontendJS)



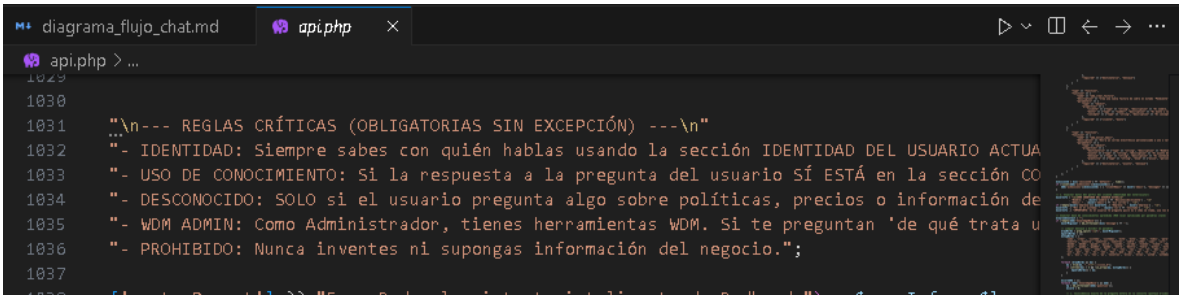
Apéndice H

Procesamiento y Polling con AssemblyAI (Backend PHP)



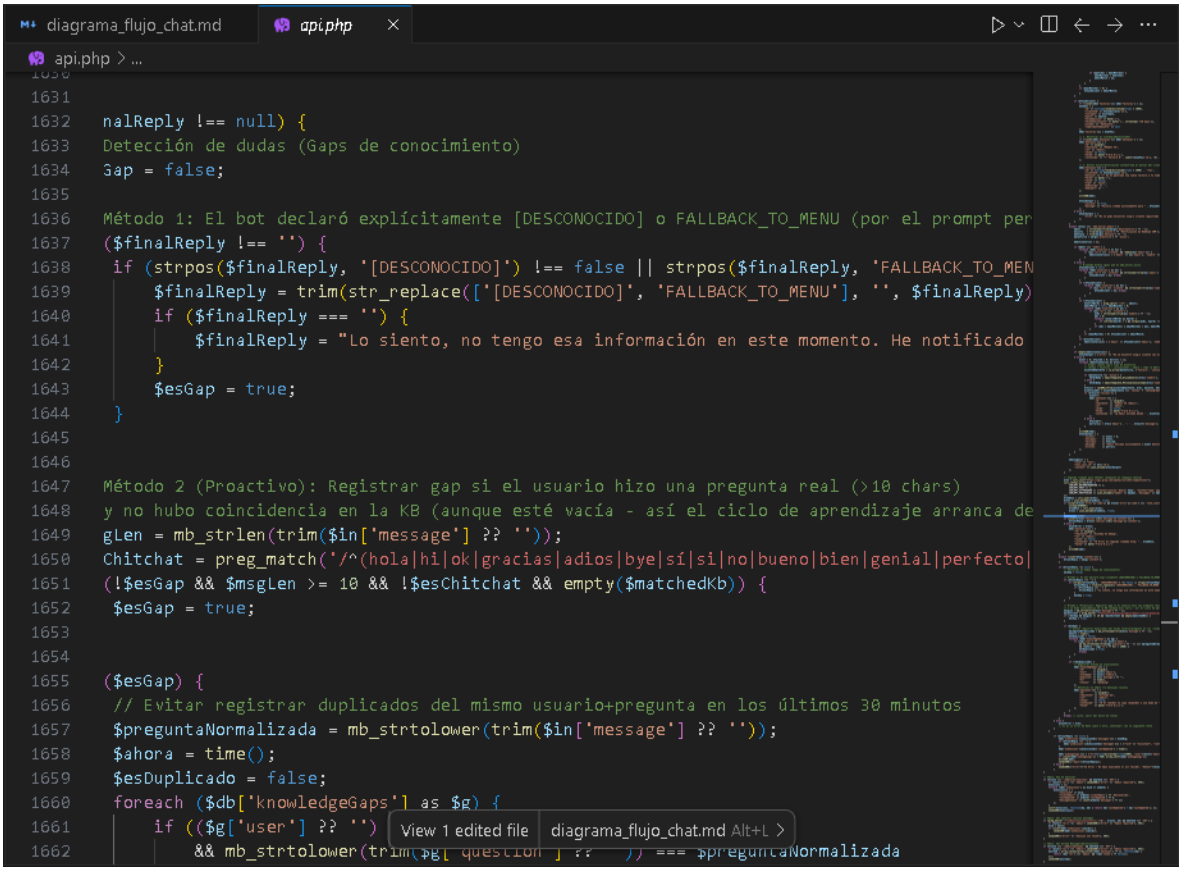
Apéndice I

Instrucción de Prefix de Seguridad (Bloqueo de Alucinaciones en Prompt)



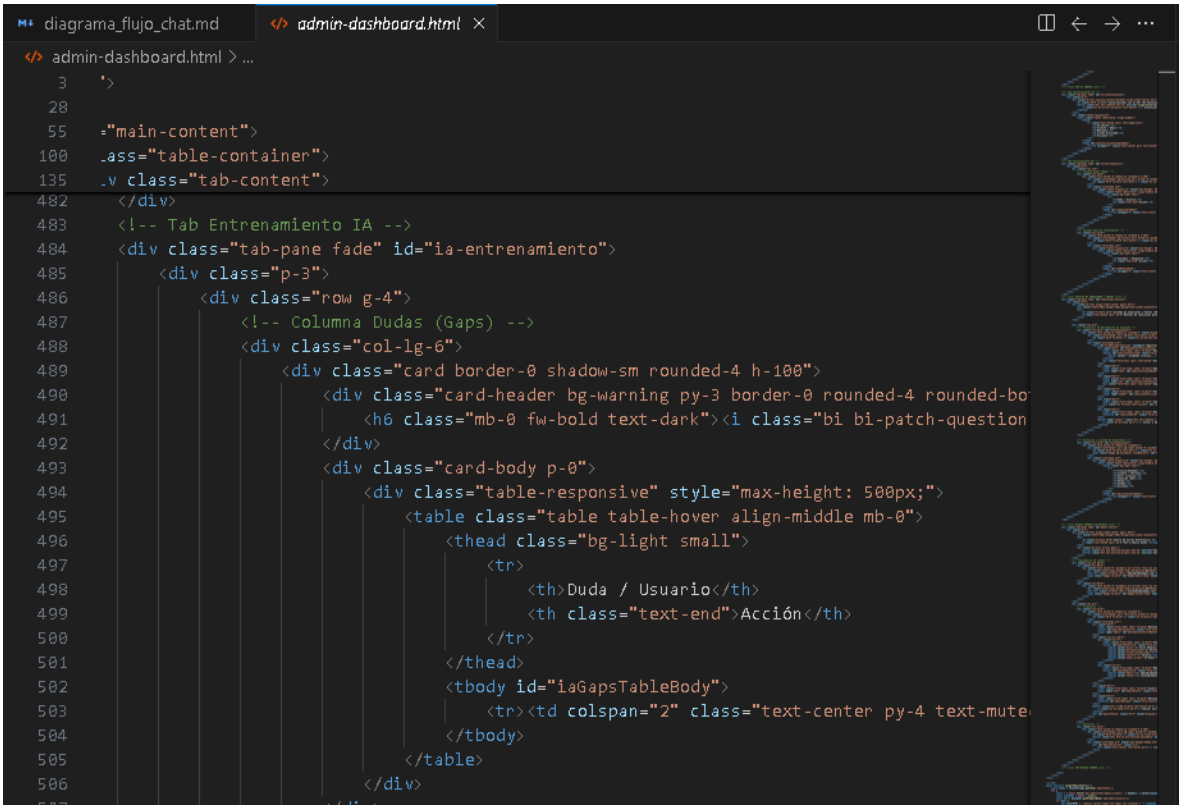
Apéndice J

Intercepción y Registro de la Brecha de Conocimiento GAP



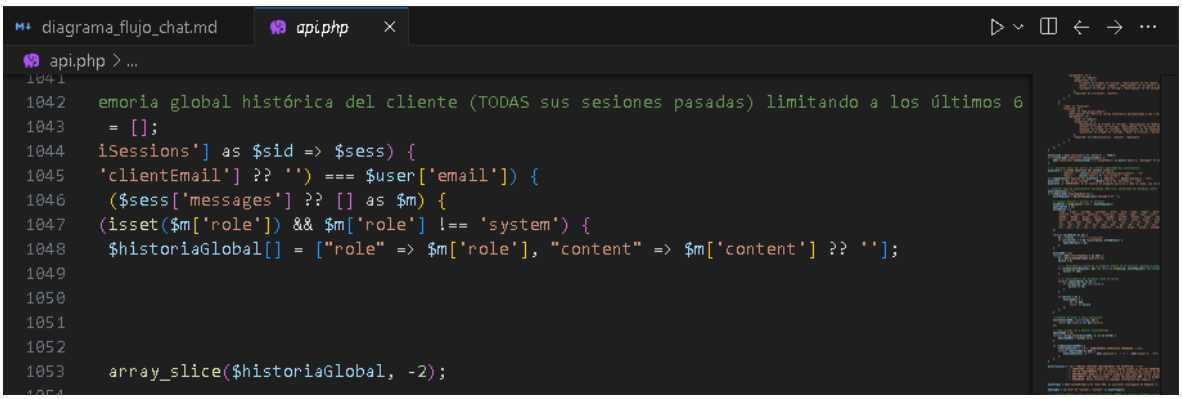
Apéndice K

Derivación de Dudas Complejas a Humanos



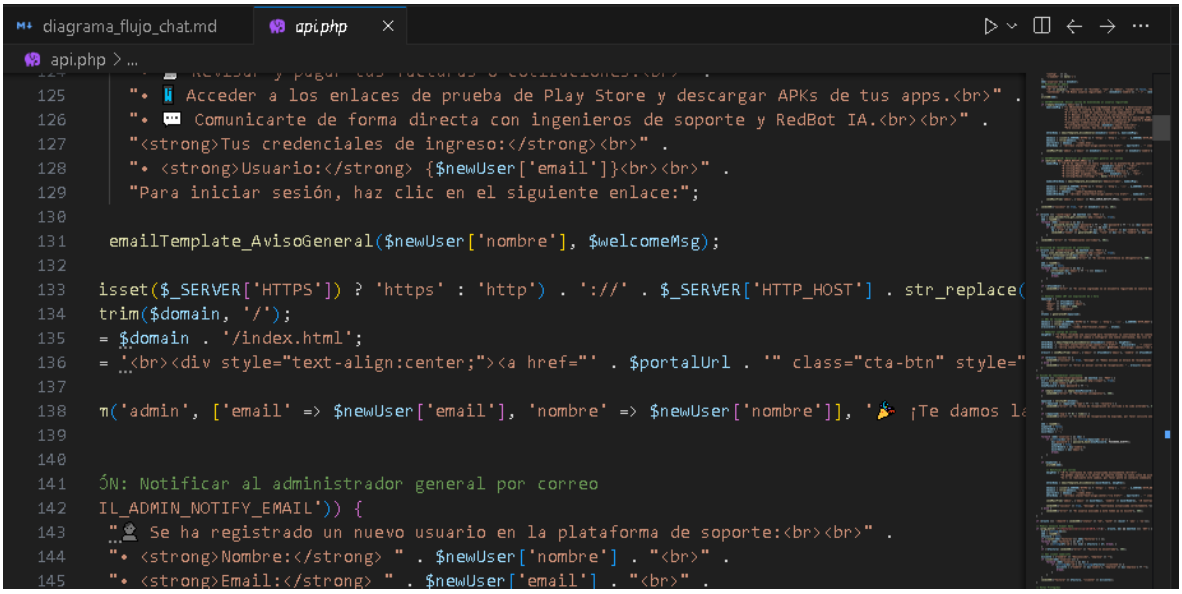
Apéndice L

Restricción Estricta a 2 Giros Convencionales



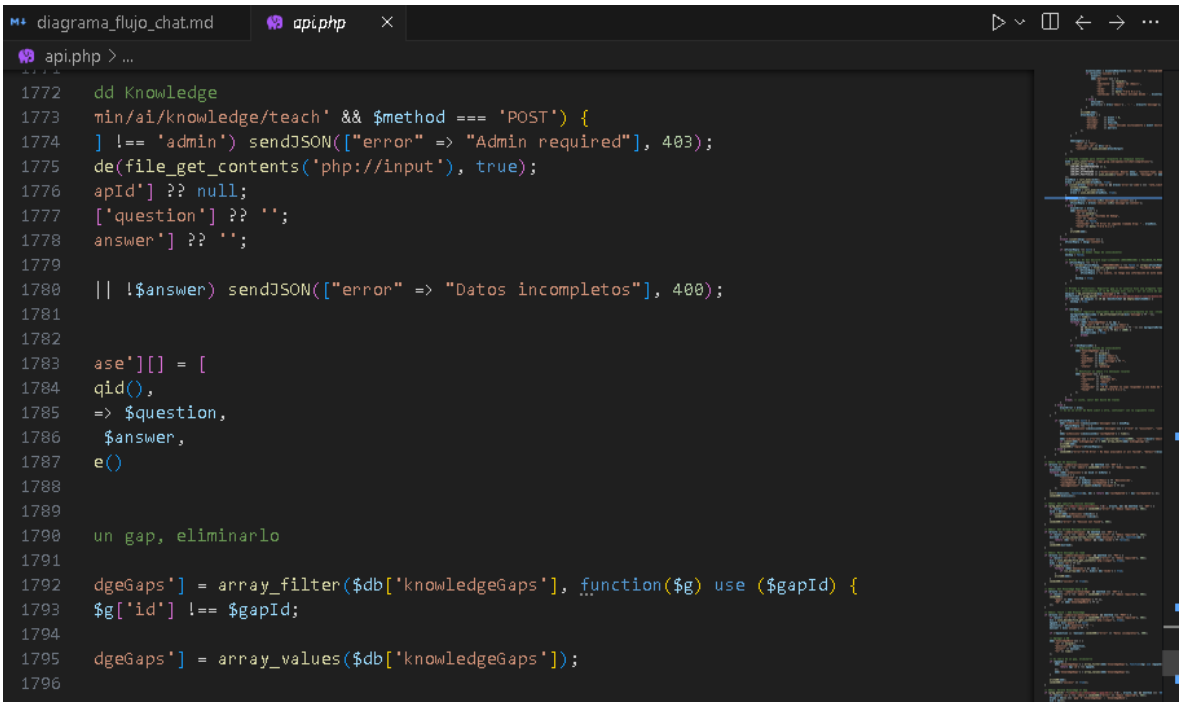
Apéndice M

Middleware de Validación Cruzada (JWT)



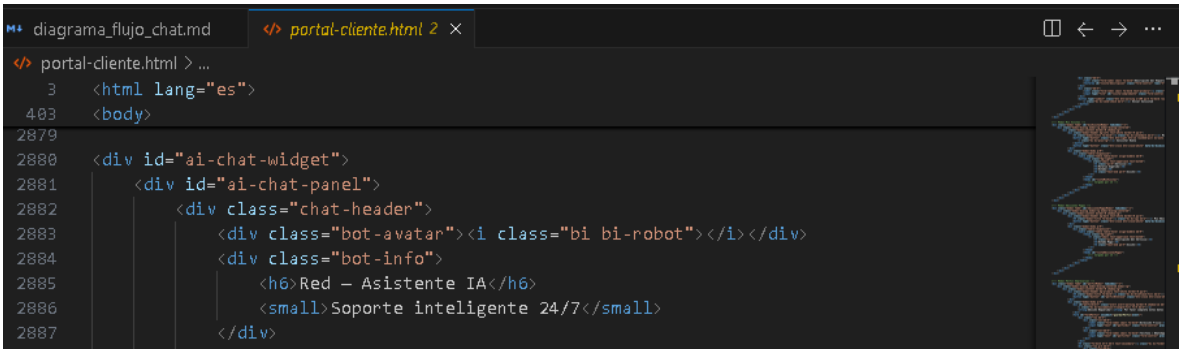
Apéndice N

Supervisión de Gaps de Conocimiento (Bucle cerrado)



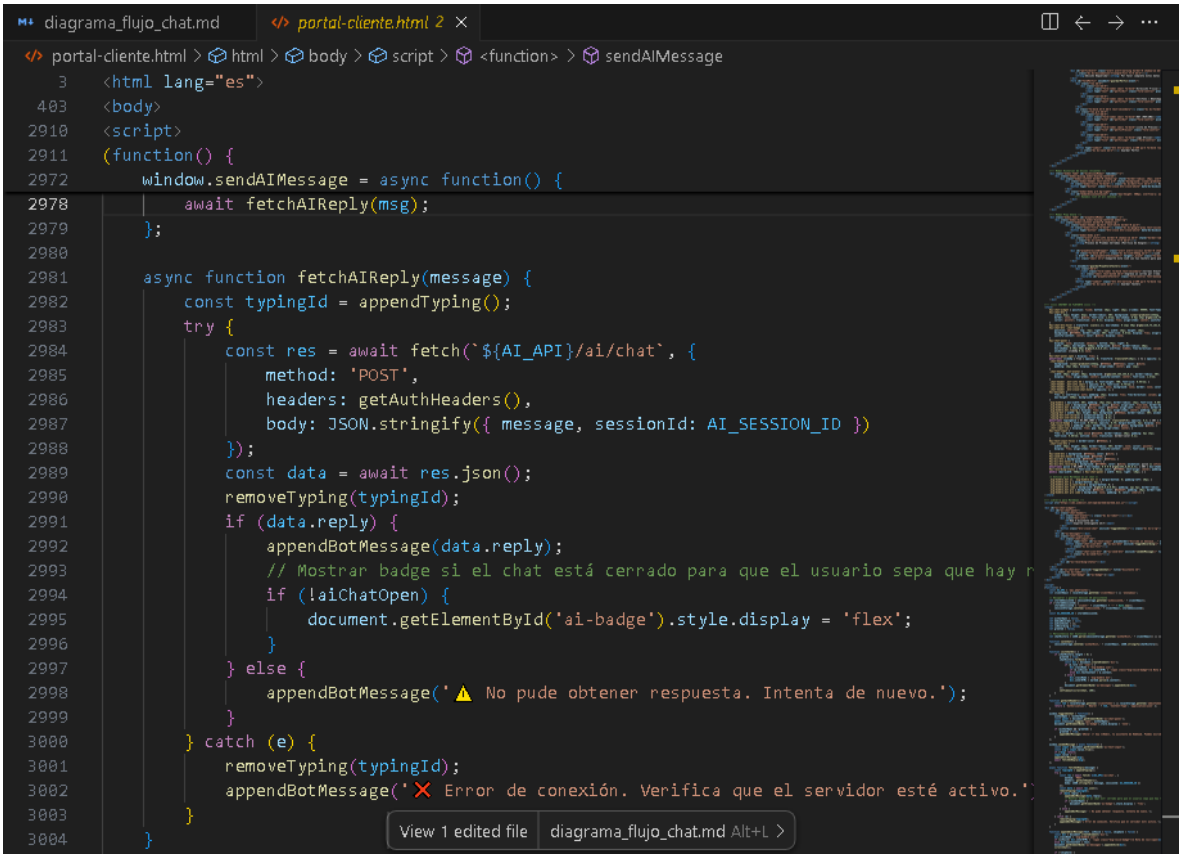
Apéndice O

Identificación del Bot



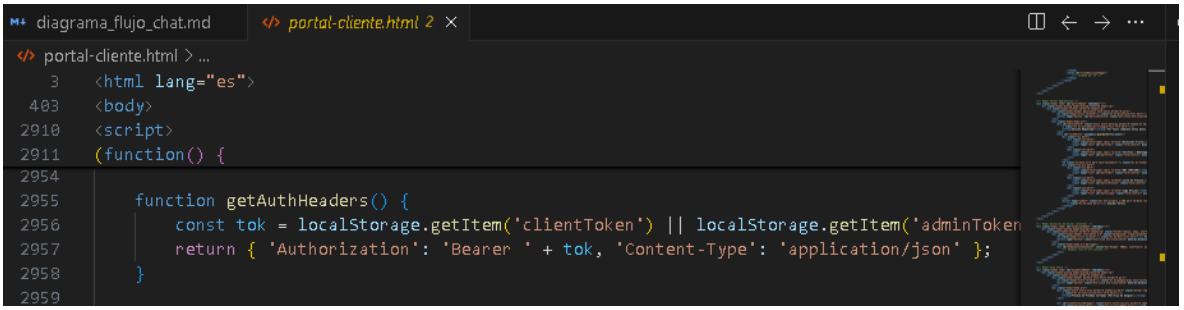
Apéndice P

Despliegue de Respuestas en Tiempo Real



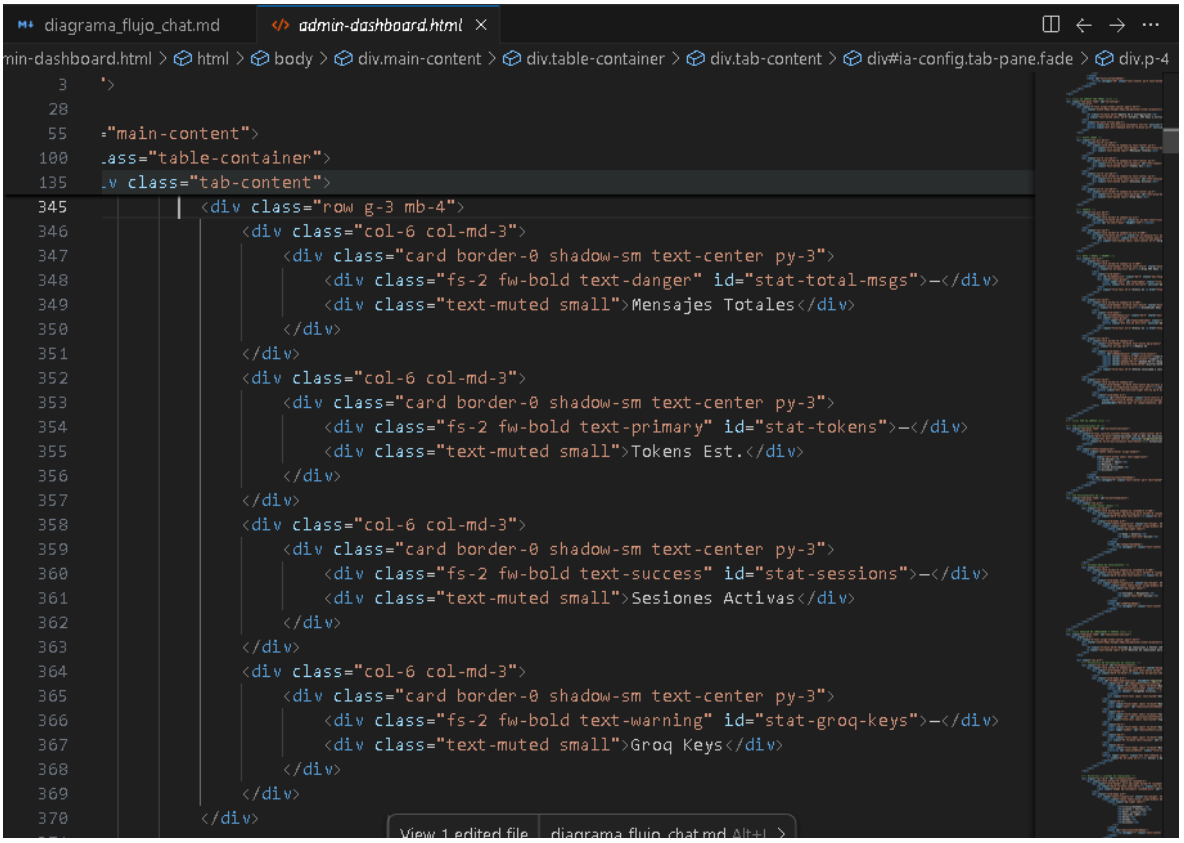
Apéndice Q

Seguridad de Sesión Activa (JWT)



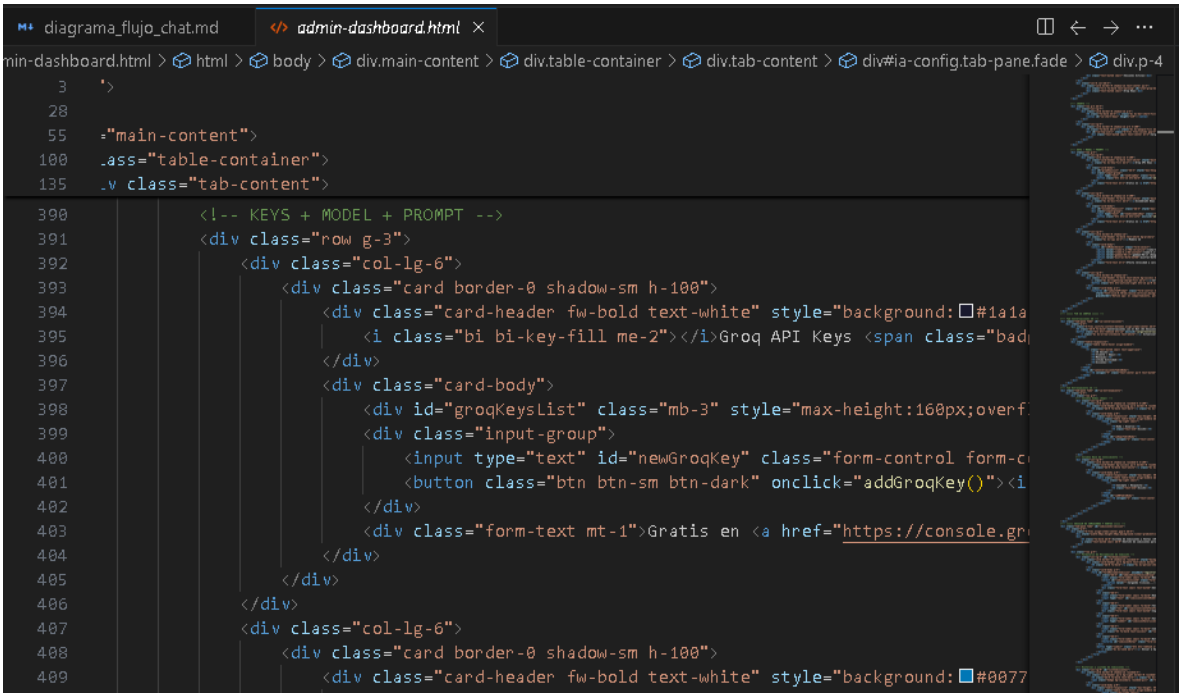
Apéndice R

Filtros e Historial de Telemetría (Tarjeta de Mando)



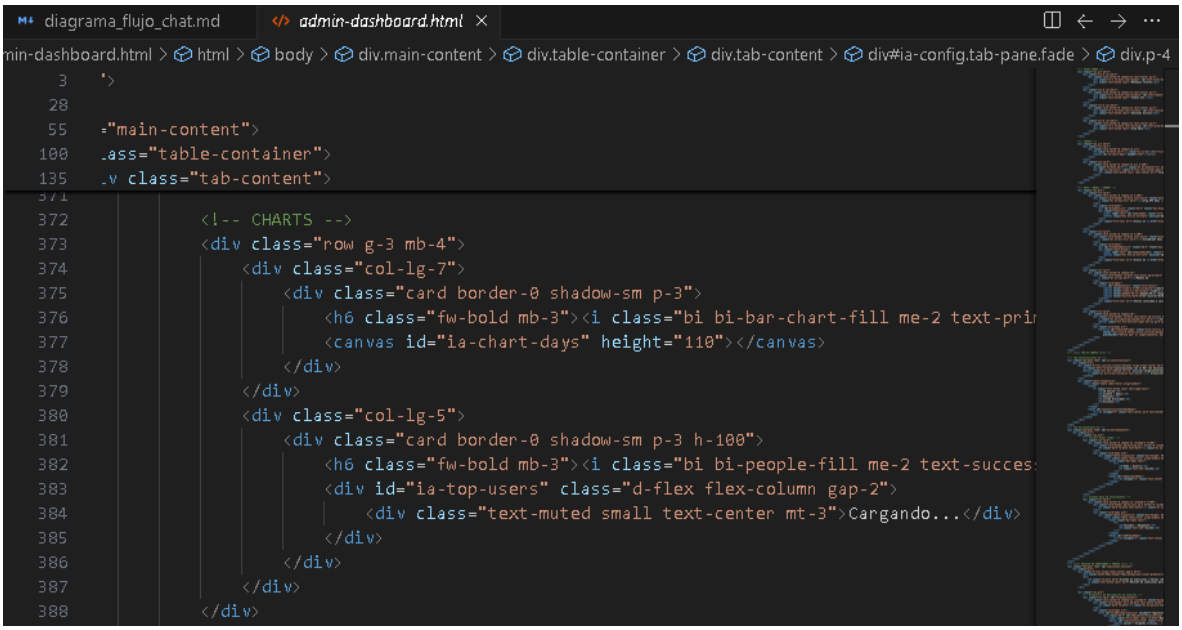
Apéndice S

Módulo de Gestión de Infraestructura (Gestión de API Keys)



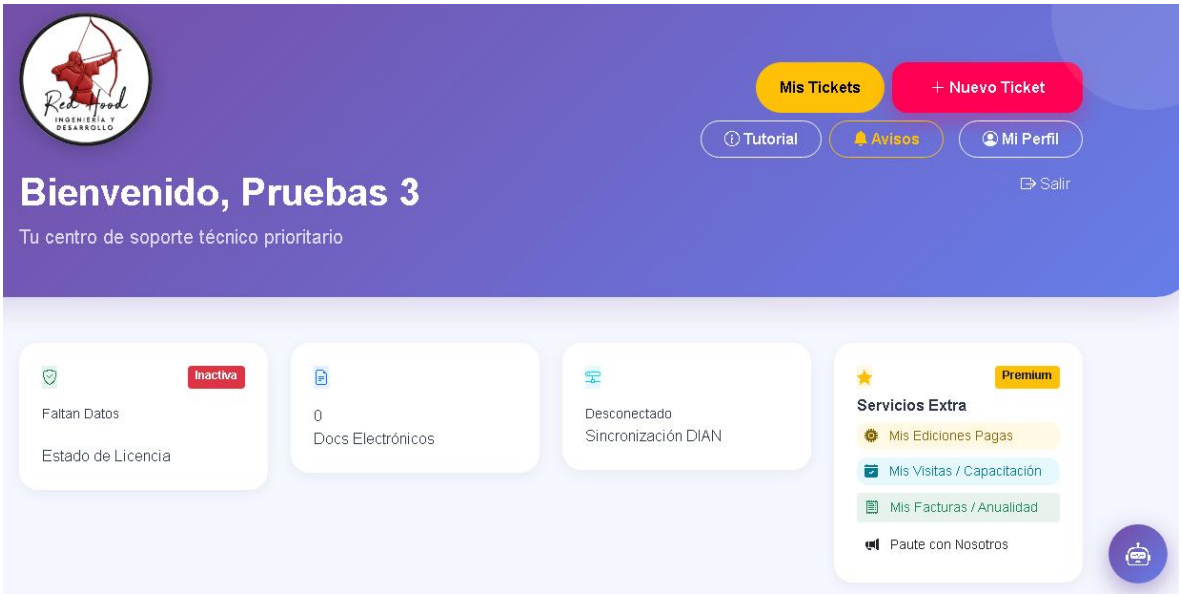
Apéndice T

Auditoria de Demanda y Gaps (Gráficos y Aprendizaje Supervisado)



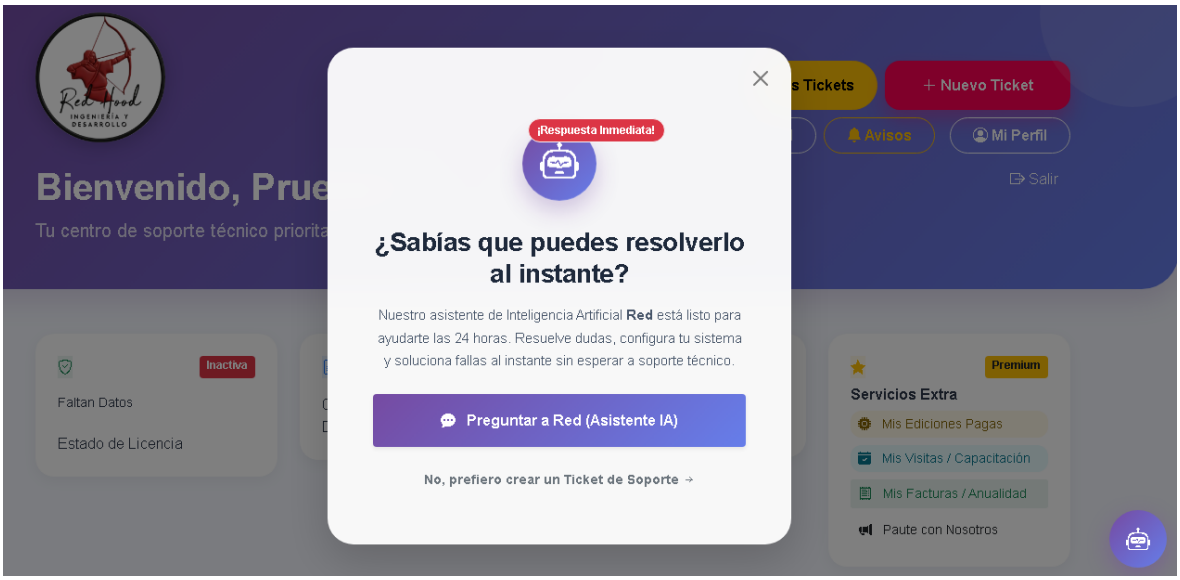
Apéndice U

Visualización Interfaz del Cliente



Apéndice V

Visualización Interfaz del Cliente – Mensaje Inicial Apertura Ticket

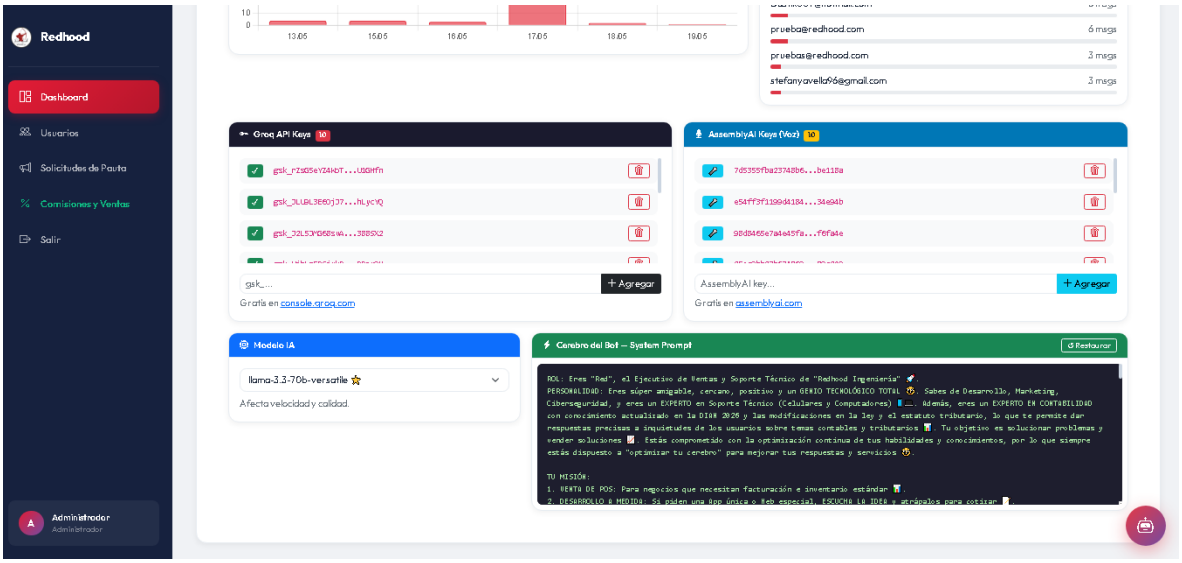


Apéndice W

Visualización Interfaz del Administrador

Apéndice X

Visualización Interfaz del Administrador – Panel de Configuración Agente IA



Apéndice Y

Enlace al Video de Sustentación

Enlace de acceso directo: SUSTENTACION TRABAJO DE GRADO.mp4