Desarrollo de un agente conversacional inteligente con embeddings para atención automatizada en WhatsApp: caso aplicado en Créalo Digital

Miguel Angel Gutierrez Maya

Director:

Julian Andres Ruiz

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas Tecnología e Ingeniería

Ingeniería de Sistemas

2026

Resumen

Este trabajo presenta el diseño, desarrollo y validación de un agente conversacional inteligente orientado a optimizar la atención al cliente en la empresa Créalo Digital. La propuesta surge ante la necesidad de reducir los tiempos de respuesta, mejorar la pertinencia de la información suministrada y disminuir la dependencia de la atención manual en el canal de WhatsApp. Para ello, se implementó una solución basada en una arquitectura de recuperación aumentada por generación (RAG), apoyada en técnicas de procesamiento de lenguaje natural, embeddings semánticos y servicios desplegados en la nube. El sistema integra una base de conocimiento empresarial, un mecanismo de recuperación contextual de información y un canal de interacción mediante WhatsApp Business API. Metodológicamente, el proyecto se desarrolló bajo un enfoque de investigación aplicada con componentes cualitativos y cuantitativos, que permitieron analizar la problemática inicial, definir los requerimientos del sistema y validar el desempeño del prototipo mediante pruebas piloto. Los resultados evidencian la viabilidad del uso de inteligencia artificial en contextos PyME, al demostrar mejoras en la oportunidad de respuesta, la consistencia de la información entregada y la eficiencia operativa del proceso de atención al cliente.

Palabras clave: *agente conversacional, embeddings, recuperación semántica, RAG, inteligencia artificial, WhatsApp Business, atención al cliente, PyME.*

Abstract

This paper presents the design, development, and validation of an intelligent conversational agent aimed at optimizing customer service processes in the company Créalo Digital. The proposal arises from the need to reduce response times, improve the relevance of the information provided, and decrease dependence on manual support through the WhatsApp channel. To address this need, a solution based on a Retrieval-Augmented Generation (RAG) architecture was implemented, supported by natural language processing techniques, semantic embeddings, and cloud-based services. The system integrates an enterprise knowledge base, a contextual information retrieval mechanism, and an interaction channel through the WhatsApp Business API. Methodologically, the project was developed under an applied research approach with qualitative and quantitative components, which made it possible to analyze the initial problem, define the system requirements, and validate the prototype's performance through pilot testing. The results demonstrate the feasibility of using artificial intelligence in SME environments by showing improvements in response timeliness, consistency of delivered information, and operational efficiency in customer service processes.

Keywords: *conversational agent, embeddings, semantic retrieval, RAG, artificial intelligence, WhatsApp Business, customer service, SME.*

Tabla de contenidos

Introducción 7

Líneas y grupos de interés investigativo 9

Planteamiento del problema 10

Justificación 13

Objetivos 15

Marco teórico 16

Marco conceptual 19

Marco legal 21

Marco tecnológico 24

Metodología de Investigación 28

Metodología de desarrollo de software 34

Resultados de métricas de validación del prototipo 54

Limitaciones, riesgos y consideraciones del modelo 57

Cronograma de actividades 59

Recursos necesarios para la implementación 63

Conclusiones 64

Anexos 66

Referencias 68

Lista de tablas

Tabla 1 Relación de intereses investigativos, líneas y grupos de investigación. 9

Tabla 2 Resultados de métricas de validación del prototipo. 55

Tabla 3 Descripción del cronograma de actividades por fases 59

Tabla 4 Relación entre las fases del cronograma y las metodologías del proyecto 62

Tabla 5 Presupuesto por recursos 63

Lista de figuras

Figura 1 Árbol de problemas del proceso actual de atención al cliente en Créalo Digital.....12

Figura 2 Estructura matemática del mecanismo de autoatención utilizado en modelos Transformer.....16

Figura 3 Interacción del usuario con el asistente conversacional a través de WhatsApp Business. . 44

Figura 4 Arquitectura orientada a microservicios del prototipo implementado en Créalo Digital.... 46

Figura 5 Panel principal de la plataforma para la supervisión de métricas y actividades del asistente. 48

Figura 6 Panel de configuración de documentos y productos para embeddings.....50

Figura 7 Panel de configuración general del asistente conversacional. 51

Introducción

En la actualidad, la inteligencia artificial se ha convertido en un componente estratégico para el fortalecimiento de los procesos empresariales, especialmente en el ámbito de la atención al cliente. Las organizaciones que logran integrar soluciones tecnológicas inteligentes son capaces de responder de manera más ágil, precisa y personalizada a las necesidades de sus usuarios, mejorando así su competitividad en entornos digitales.

En el contexto colombiano, el crecimiento sostenido del comercio electrónico ha impulsado la necesidad de automatizar canales de atención, siendo WhatsApp una de las plataformas más utilizadas para la comunicación comercial. Según la Cámara Colombiana de Comercio Electrónico (2024), las ventas en línea alcanzaron los 62,1 billones de pesos en 2023, lo que refleja la creciente demanda de herramientas tecnológicas que soporten la interacción con los clientes de forma eficiente.

En este escenario, la empresa Créalo Digital, dedicada al desarrollo de soluciones tecnológicas y estrategias de marketing digital, ha identificado la necesidad de implementar un sistema que permita responder de manera automática y coherente a las consultas de sus clientes. Para ello, se propone el desarrollo de un agente conversacional inteligente, sustentado en modelos de embeddings y arquitecturas Transformer, capaces de comprender el lenguaje natural y recuperar información relevante a partir de la base de conocimiento de la empresa.

El presente trabajo de investigación aplicada combina la rigurosidad científica con el desarrollo tecnológico, articulando dos componentes fundamentales: la metodología

investigativa, enfocada en la comprensión del problema y la validación del prototipo, y la metodología de desarrollo, orientada a la construcción y prueba del sistema.

De esta manera, este proyecto busca contribuir no solo a la mejora de los procesos de atención en Créalo Digital, sino también al avance del conocimiento en el campo del procesamiento de lenguaje natural y la inteligencia artificial aplicada al sector empresarial colombiano.

Líneas y grupos de interés investigativo

Tabla 1

Relación de intereses investigativos, líneas y grupos de investigación.

Intereses en ingeniería e investigación	Línea de investigación y áreas temáticas	Grupo de investigación
Proyecto aplicado	Ingeniería del software	No aplica

Nota. Elaboración propia a partir de la clasificación del proyecto dentro de la línea de ingeniería del software.

Planteamiento del problema

La comunicación digital se ha convertido en un pilar estratégico para las empresas que buscan mantener relaciones sólidas con sus clientes. En Colombia, el comercio electrónico ha mostrado un crecimiento sostenido en los últimos años: en 2023 se registraron más de 370,5 millones de transacciones en ventas en línea, con un incremento del 11,5 % respecto al año anterior (Forbes Colombia, 2024). A su vez, la Cámara Colombiana de Comercio Electrónico (2024) reportó que el valor total de ventas alcanzó los 62,1 billones de pesos, reflejando un aumento del 12,58 % frente al 2022. Este crecimiento demuestra que los canales digitales, particularmente aplicaciones de mensajería como WhatsApp, son fundamentales para la interacción comercial. Sin embargo, muchas pequeñas y medianas empresas (pymes) enfrentan limitaciones en la atención a clientes: tiempos de respuesta prolongados, respuestas poco personalizadas y dificultades para gestionar información actualizada sobre productos o servicios.

En el caso de Créalo Digital, empresa de soluciones tecnológicas y de mercadeo digital, se ha identificado la necesidad de mejorar los tiempos de respuesta y la pertinencia de la información entregada a los clientes. Actualmente, la atención depende de la disponibilidad de personal humano, lo que genera demoras, pérdida de oportunidades comerciales y disminución en la satisfacción del cliente.

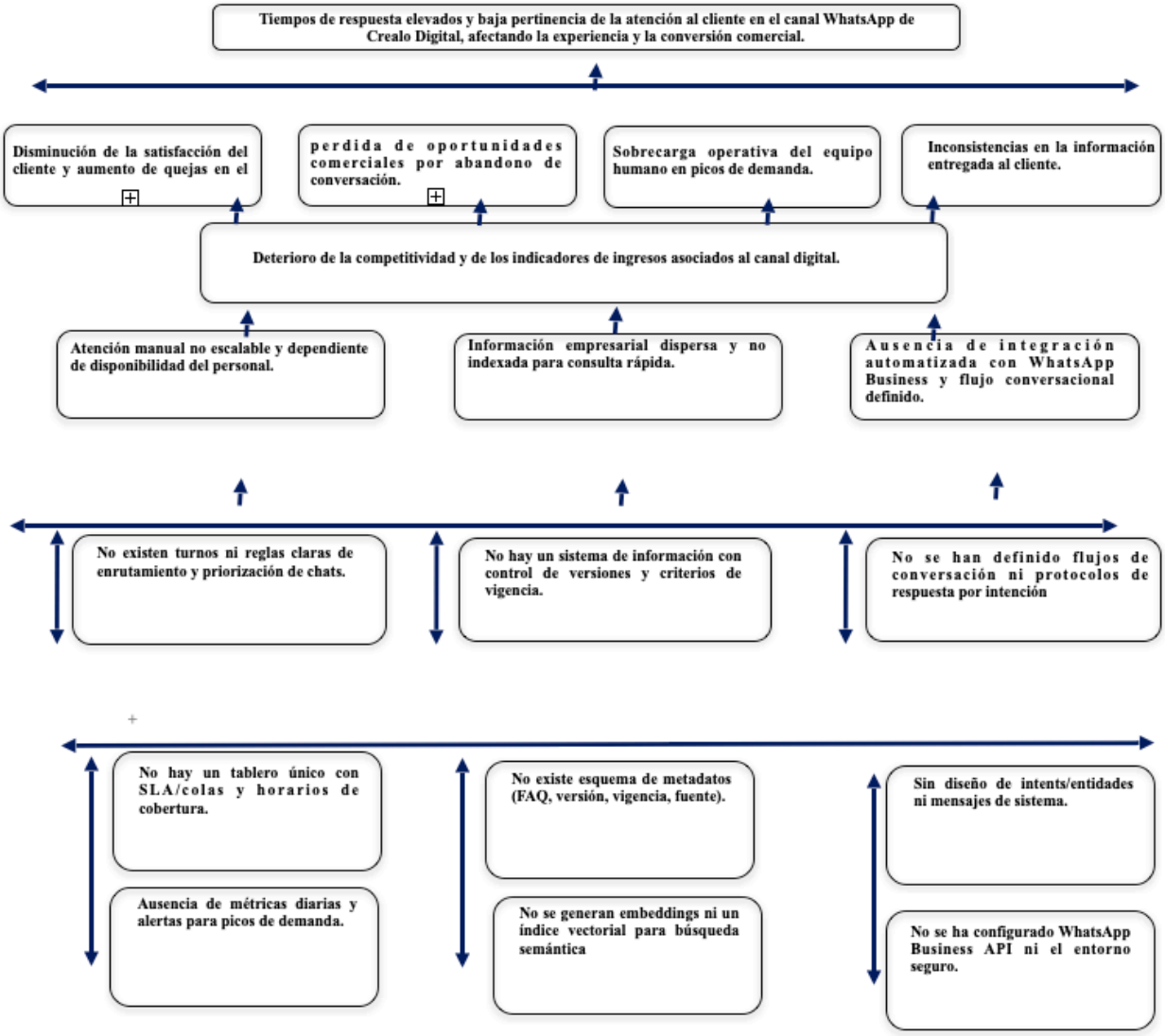
La pregunta central que orienta este proyecto es:

¿Cómo desarrollar un agente conversacional inteligente que, mediante técnicas de inteligencia artificial y uso de embeddings, permita a Créalo Digital optimizar la atención al cliente en WhatsApp, garantizando respuestas rápidas, pertinentes y basadas en información actualizada de sus servicios y productos?

La magnitud del problema no solo se refleja en la pérdida de eficiencia, sino también en el impacto en la percepción de calidad del servicio, aspecto crítico para empresas en el sector digital. El reto está en combinar metodologías de procesamiento de lenguaje natural con herramientas accesibles para una empresa en crecimiento, logrando un equilibrio entre innovación tecnológica y aplicabilidad práctica.

Figura 1

Árbol de problemas del proceso actual de atención al cliente en Créalo Digital.



Nota. Esta figura muestra la relación entre causas, problema central y consecuencias asociadas al modelo de atención previo a la implementación del asistente conversacional.

Justificación

La pertinencia de este proyecto se sustenta en tres dimensiones:

1. Académica y disciplinar:

El proyecto permite aplicar conocimientos de inteligencia artificial, procesamiento de lenguaje natural y desarrollo de sistemas de información en un caso real. Además, introduce a la práctica empresarial los fundamentos de arquitecturas modernas como Transformers y el uso de embeddings para la recuperación semántica de información, temas ampliamente estudiados en la literatura científica (Vaswani et al., 2017; Devlin et al., 2019). Su desarrollo aporta evidencia sobre la aplicabilidad de conceptos avanzados de IA en contextos empresariales locales.

2. Social:

Mejorar la atención al cliente mediante un agente conversacional inteligente contribuye a democratizar el acceso a tecnologías emergentes en las pymes colombianas. Un sistema de este tipo permite a los clientes obtener información clara, inmediata y confiable, reduciendo la brecha entre las grandes empresas que ya implementan este tipo de soluciones y las medianas empresas que buscan competir en el mercado digital.

3. Empresarial y personal:

Para Créalo Digital, el proyecto representa una oportunidad de innovación tecnológica que fortalece su posicionamiento competitivo, optimiza recursos y mejora la satisfacción del cliente. Para los investigadores y estudiantes involucrados, significa la posibilidad de adquirir experiencia en el diseño e implementación de soluciones con impacto real, integrando teoría y práctica en un escenario empresarial.

En conclusión, este proyecto se justifica porque responde a una necesidad concreta del sector empresarial, se alinea con tendencias globales de automatización e inteligencia artificial, y contribuye al desarrollo académico y profesional en el área de la ingeniería de sistemas y el desarrollo de software aplicado.

Objetivos

Objetivo general

Desarrollar un agente conversacional inteligente integrado a WhatsApp Business que, mediante el uso de técnicas de embeddings y procesamiento de lenguaje natural, permita a la empresa Créalo Digital optimizar la atención al cliente ofreciendo respuestas rápidas, pertinentes y basadas en información actualizada de sus productos y servicios.

Objetivos específicos

Diseñar la arquitectura del sistema de información que permita indexar, almacenar y actualizar el conocimiento empresarial utilizando métodos de embeddings para la recuperación semántica de información.

Implementar un agente conversacional basado en inteligencia artificial, integrado con la API de WhatsApp Business, que gestione interacciones automáticas con clientes de manera contextual y personalizada.

Validar el desempeño del agente conversacional mediante pruebas piloto con clientes reales de Créalo Digital, evaluando métricas de satisfacción, pertinencia de las respuestas y tiempos de respuesta frente al modelo actual de atención.

Documentar el proceso de desarrollo, implementación y validación del prototipo, generando evidencia académica y técnica que aporte al conocimiento en el campo de los agentes conversacionales inteligentes aplicados a pymes en el contexto colombiano.

Marco teórico

El marco teórico de este proyecto se apoya en la literatura sobre procesamiento de lenguaje natural, arquitecturas de atención y sistemas de información empresarial:

Transformers y el modelo Attention is All You Need:

Vaswani et al. (2017) introdujeron el modelo Transformer, cuya base es el mecanismo de auto-atención (self-attention), que calcula relaciones de dependencia entre todas las palabras de una secuencia de manera paralela. En la figura 2 se puede evidenciar matemáticamente como se expresa el modelo de atención.

Figura 2

Estructura matemática del mecanismo de autoatención utilizado en modelos Transformer.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Nota. Operación interna del mecanismo de self-attention, base teórica para la construcción del sistema de recuperación semántica del asistente conversacional.

Modelos de embeddings y recuperación de información:

Desde Word2Vec (Mikolov et al., 2013) hasta modelos más avanzados como BERT (Devlin et al., 2019), los embeddings han evolucionado para capturar relaciones semánticas complejas. En este proyecto, los embeddings se aplicarán a la indexación de documentos internos de Créalo Digital, lo cual permitirá responder preguntas de clientes con información precisa y actualizada.

Aplicaciones empresariales de chatbots con IA:

Investigaciones recientes (Cui et al., 2020; Adamopoulou & Moussiades, 2020) muestran que la implementación de chatbots inteligentes en plataformas de mensajería mejora la disponibilidad del servicio al cliente, incrementa la retención y optimiza procesos de soporte.

Sistemas de información para la gestión del conocimiento:

El proyecto integra el concepto de sistemas de información orientados a la gestión del conocimiento, en los cuales la información se organiza, indexa y hace accesible para los usuarios a través de interfaces inteligentes. Esto permite cerrar la brecha entre el conocimiento interno de la empresa y las necesidades de los clientes.

En el contexto internacional, diversas pequeñas y medianas empresas (PyMEs) han comenzado a incorporar soluciones basadas en inteligencia artificial conversacional como estrategia para mejorar la atención al cliente. Proyectos como “Tiledesk AI Chatbot for SMEs” (Italia, 2023) y “ChatGPT Plugin for Shopify Stores” (EE. UU., 2024) demostraron que los

agentes conversacionales permiten reducir en un 40 % los tiempos de respuesta y aumentar la retención de clientes en canales digitales.

En Latinoamérica, casos como el de “Aivo” (Argentina, 2022) y “Yellow Messenger” (Brasil, 2023) han logrado integrar sistemas de Retrieval-Augmented Generation (RAG) en entornos de e-commerce y servicios financieros, combinando bases de conocimiento con modelos de lenguaje de gran escala. Estas experiencias evidencian la viabilidad de adoptar tecnologías de inteligencia artificial en contextos empresariales no masivos.

En Colombia, se identifican iniciativas como “PymeBot” (2022), desarrollada en colaboración con la Cámara de Comercio de Bogotá, que aplicó modelos de PLN para automatizar la atención a empresarios y mejorar el acceso a servicios digitales. Sin embargo, la mayoría de estos proyectos se basan en modelos de flujo predefinido y carecen de indexación semántica mediante embeddings.

En ese sentido, el presente proyecto se diferencia por su enfoque en la recuperación contextual (RAG), que combina técnicas de embeddings y búsqueda vectorial para ofrecer respuestas dinámicas y contextualizadas a partir de la información empresarial. Esta innovación refuerza la pertinencia del trabajo en el contexto nacional e internacional de las PyMEs tecnológicas.

Marco conceptual

Agentes conversacionales (Chatbots):

Son sistemas de software diseñados para interactuar con usuarios mediante lenguaje natural. A diferencia de los chatbots basados en reglas, los agentes inteligentes integran modelos de procesamiento de lenguaje natural (PLN) que permiten comprender la intención del usuario y responder de forma contextual (Jurafsky & Martin, 2023).

WhatsApp Business API:

Herramienta oficial de Meta que permite a las empresas automatizar y gestionar interacciones con clientes a través de WhatsApp. Provee un canal masivo y estandarizado para integrar agentes conversacionales en un entorno de alta adopción.

Embeddings:

Representaciones vectoriales de palabras, frases o documentos que capturan su significado semántico. Permiten medir similitudes entre textos en un espacio de alta dimensión, facilitando la recuperación de información relevante (Mikolov et al., 2013). En este proyecto se utilizarán embeddings para indexar información de la empresa Créalo Digital y responder preguntas de los clientes.

Recuperación semántica de información:

Proceso mediante el cual se consulta una base de datos de embeddings y se obtiene la información más similar al requerimiento del usuario. Es clave en agentes conversacionales que necesitan responder basados en datos empresariales actualizados.

Inteligencia Artificial aplicada a la atención al cliente:

La IA permite optimizar tiempos de respuesta, reducir costos operativos y mejorar la experiencia del cliente, generando un impacto positivo tanto en la productividad de la empresa como en la satisfacción del usuario.

Marco legal

El desarrollo de un agente conversacional inteligente para la atención de clientes en WhatsApp Business debe enmarcarse en el cumplimiento de las disposiciones legales vigentes en Colombia, así como en lineamientos internacionales relacionados con la protección de datos, el comercio electrónico y el uso de plataformas tecnológicas. El marco jurídico del proyecto se estructura en los siguientes aspectos:

1. Protección de datos personales y Habeas Data

En Colombia, la Ley 1581 de 2012 establece las disposiciones generales para la protección de datos personales. Esta normativa regula la recolección, almacenamiento, uso y circulación de información de carácter privado, garantizando los derechos de los titulares sobre el acceso, corrección y supresión de sus datos. En el contexto de este proyecto, la implementación del chatbot debe ajustarse a los principios de finalidad, libertad, veracidad, seguridad y confidencialidad, asegurando que la información de los clientes sea utilizada exclusivamente para los fines autorizados. Complementariamente, el Decreto 1377 de 2013 reglamenta la Ley 1581 y refuerza la necesidad de contar con mecanismos claros de autorización informada para el tratamiento de datos, aspecto que se aplicará en el consentimiento otorgado por los clientes al interactuar con el agente conversacional.

2. **Regulación de comercio electrónico**

   La Ley 527 de 1999 define y reglamenta el uso de mensajes de datos, firmas digitales y documentos electrónicos en Colombia. Este marco legal otorga validez jurídica a las interacciones y transacciones realizadas mediante medios electrónicos. Si bien el chatbot no procesará pagos en su fase inicial, sí generará registros de conversaciones que constituyen mensajes de datos y, por tanto, deberán ser tratados bajo los principios de integridad y autenticidad establecidos en la ley.

3. **Normatividad internacional en privacidad y uso de plataformas**

   El uso de servicios provistos por Meta a través de la WhatsApp Business API está regulado por los Términos de Servicio y Políticas de Uso de Meta, que exigen a las empresas cumplir con la normativa local de protección de datos y garantizar la transparencia en el manejo de la información de los usuarios. Asimismo, la normativa internacional, como el Reglamento General de Protección de Datos (GDPR, 2016/679) de la Unión Europea, aunque no aplica de manera obligatoria en Colombia, constituye una referencia relevante para la adopción de buenas prácticas en la gestión de datos sensibles. Tomar en cuenta principios como minimización de datos, consentimiento explícito y derecho al olvido permitirá fortalecer la confianza en el sistema y anticipar posibles regulaciones futuras.

4. Propiedad intelectual y confidencialidad

De acuerdo con la Decisión Andina 486 de 2000 sobre propiedad industrial y la legislación colombiana en materia de derechos de autor, el desarrollo del software generado en este proyecto se considerará una obra protegida. En este sentido, los resultados deben contemplar acuerdos de confidencialidad entre los estudiantes investigadores, la universidad y la empresa Créalo Digital, con el fin de salvaguardar tanto la información empresarial como los derechos patrimoniales y morales de los autores.

Marco tecnológico

El desarrollo de un agente conversacional inteligente requiere la integración de tecnologías de software, arquitecturas de inteligencia artificial y herramientas de comunicación que garanticen la eficiencia, escalabilidad y seguridad del sistema. Este proyecto combina recursos de desarrollo de software modernos con componentes de inteligencia artificial orientados al procesamiento del lenguaje natural (PLN), junto con los servicios de comunicación provistos por Meta a través de la API de WhatsApp Business.

Lenguajes y entornos de programación

El proyecto se implementará utilizando Node.js, seleccionado por su versatilidad, soporte comunitario y compatibilidad con bibliotecas de inteligencia artificial y servicios web. Node.js se aplicará en la construcción de servicios backend y en la integración con la API de WhatsApp Business, permitiendo la gestión de mensajes, sesiones de usuarios y solicitudes en tiempo real. Su arquitectura basada en eventos es ideal para manejar múltiples conexiones simultáneas de forma eficiente.

Frameworks y bibliotecas de IA

La implementación de los modelos semánticos y de embeddings se apoyará en librerías reconocidas internacionalmente:

OpenAI: proporciona modelos de lenguaje pre-entrenados como text-embeddings-small, que permiten realizar tareas de comprensión y recuperación de información contextual.

Bases de datos y almacenamiento de conocimiento

El agente conversacional contará con un sistema de información interno encargado de almacenar y gestionar la base de conocimiento de la empresa.

Se utilizará PostgreSQL como base de datos relacional principal, ideal para manejar datos estructurados y consultas complejas.

De manera complementaria, MongoDB puede emplearse para almacenar conversaciones y registros en formato JSON, ofreciendo flexibilidad en la gestión de datos semiestructurados. Para la búsqueda semántica, se implementará un índice vectorial que permita comparar las representaciones de texto (embeddings) y recuperar la información más relevante a partir de las consultas del usuario.

Integración con WhatsApp Business API

La comunicación con los usuarios se gestionará a través de la WhatsApp Business API, herramienta oficial de Meta, que permite la automatización de conversaciones en el entorno empresarial.

El bot actuará como un intermediario entre el cliente y el sistema de información, procesando los mensajes recibidos, interpretando su intención mediante el modelo de IA y respondiendo con información actualizada desde la base de conocimiento.

La integración se realizará bajo los lineamientos técnicos y jurídicos de Meta, garantizando el cumplimiento de las políticas de uso y privacidad de datos.

Infraestructura y servicios en la nube

El despliegue del prototipo se realizará utilizando servicios en la nube para asegurar disponibilidad y escalabilidad.

Amazon Web Services (AWS) será la opción principal para hospedar el backend, supabase para el almacenamiento de base de datos relacional y MongoDB atlas para el almacenamiento NoSQL.

Se emplearán servicios como AWS Lambda para la ejecución de servicios, y S3 o Cloud Storage para el almacenamiento de datos y respaldos.

La arquitectura estará diseñada bajo un modelo cliente-servidor, asegurando modularidad, mantenimiento y facilidad de actualización.

Seguridad y control de acceso

El proyecto implementará políticas de seguridad que garanticen la integridad de la información y la confidencialidad de los datos:

- Cifrado de datos en tránsito mediante HTTPS y SSL/TLS.
- Autenticación basada en tokens (JWT) para las conexiones entre servicios.
- Control de roles y permisos para la administración del sistema de información.
- Cumplimiento con los principios de la Ley 1581 de 2012 sobre protección de datos personales y las políticas de seguridad digital recomendadas por Meta.

Entornos de pruebas y despliegue continuo

Se establecerán entornos diferenciados para el desarrollo, pruebas y despliegue, utilizando herramientas de control de versiones y automatización:

- Git y GitHub para la gestión del código fuente y control de versiones.
- Docker para la contenedorización de los servicios, asegurando la portabilidad del sistema.
- GitHub Actions para la integración y entrega continua (CI/CD).

Metodología de Investigación

El proyecto se enmarca dentro de una investigación aplicada con un enfoque mixto (cualitativo y cuantitativo), orientada a resolver un problema real mediante la construcción y validación de un prototipo tecnológico. La investigación combina el análisis teórico con la experimentación práctica, articulando métodos de recolección de datos empíricos y su posterior análisis estadístico y descriptivo.

1. Tipo y enfoque de investigación

El enfoque mixto permite integrar la comprensión profunda del fenómeno (mediante técnicas cualitativas) con la medición objetiva del impacto (a través de instrumentos cuantitativos).

Desde la dimensión cualitativa, se busca identificar las necesidades, expectativas y problemáticas que enfrenta la empresa Créalo Digital en la atención al cliente.

Desde la dimensión cuantitativa, se medirá la eficiencia del prototipo a través de métricas verificables como el tiempo promedio de respuesta, la precisión de las respuestas y el nivel de satisfacción del cliente.

Este enfoque permite validar la hipótesis de que un agente conversacional basado en embeddings y modelos de lenguaje puede optimizar los procesos de atención al cliente en una empresa de servicios digitales.

2. Muestra y población del proyecto

A. Población objetivo

El presente estudio emplea un muestreo no probabilístico por conveniencia, fundamentado en las características propias de la población objetivo y en la naturaleza aplicada del proyecto. La empresa Créalo Digital cuenta con un grupo reducido pero representativo de clientes que mantienen interacción activa a través del canal de WhatsApp, lo cual permite capturar comportamientos y percepciones reales frente al prototipo.

Este tipo de muestreo resulta pertinente en estudios exploratorios y de validación tecnológica, donde el propósito no es la generalización estadística, sino la evaluación funcional y empírica del sistema en contextos de uso reales (Hernández-Sampieri & Mendoza Torres, 2023).

B. Muestra

La muestra definida de 20 a 30 participantes garantiza la obtención de un número suficiente de observaciones para el cálculo de métricas cuantitativas como la precisión de las respuestas y la satisfacción del usuario, mientras mantiene la viabilidad operativa y temporal del proyecto. Este enfoque se respalda en investigaciones similares de validación de chatbots y RAG en PyMEs, donde el rango de usuarios evaluadores oscila entre 15 y 40 casos (Ke et al., 2024; Seyi-Lande & Onaolapo, 2024).

De este modo, la representatividad se entiende desde una perspectiva funcional, en

la que la muestra seleccionada refleja los principales perfiles de usuarios finales del sistema, garantizando que las conclusiones obtenidas sean pertinentes y aplicables a la realidad operacional de Créalo Digital.

3. Construcción e implementación del instrumento de medición y recolección de datos

Para garantizar la validez y confiabilidad de los resultados, se emplearán diferentes instrumentos de medición, diseñados específicamente para cada etapa del proyecto:

A. Entrevistas semiestructuradas

- Objetivo: identificar necesidades, expectativas y dificultades en la atención al cliente.
- Participantes: personal administrativo y técnico de Créalo Digital.
- Aplicación: presencial o virtual durante la fase inicial del proyecto.
- Análisis: categorización de respuestas mediante análisis de contenido.

B. Encuesta de satisfacción al cliente

- Objetivo: medir la percepción del usuario sobre la calidad del servicio automatizado.
- Dimensiones: facilidad de uso, precisión de las respuestas, tiempo de respuesta y satisfacción general.
- Instrumento: cuestionario de escala Likert de 1 a 5 puntos.
- Aplicación: al finalizar la interacción del usuario con el chatbot en el piloto.

C. Registro automatizado de interacciones

- Objetivo: obtener métricas cuantitativas sobre el desempeño del agente conversacional.
- Variables: número de interacciones válidas, tiempo promedio de respuesta, tasa de error, satisfacción (por emoji o respuesta breve).
- Fuente: logs del sistema y la API de WhatsApp Business.

La encuesta tipo Likert fue sometida a validación de contenido mediante revisión por expertos en experiencia de usuario, ingeniería de software e inteligencia artificial aplicada, quienes evaluaron claridad, pertinencia y coherencia de los ítems. Posteriormente, se realizó una prueba piloto para verificar consistencia y comprensión antes de su aplicación definitiva.

4. Análisis y diagnóstico del proceso investigativo

El análisis de la información se desarrollará en tres etapas complementarias:

A. Análisis cualitativo

Se empleará el análisis de contenido para interpretar los datos obtenidos en las entrevistas con el personal de Créalo Digital.

El objetivo será identificar patrones, necesidades y oportunidades de mejora en los procesos de atención, que servirán como base para el diseño funcional del prototipo.

B. Análisis cuantitativo

Los datos obtenidos de las encuestas y registros automáticos serán procesados mediante estadística descriptiva, utilizando medidas de tendencia central (media, mediana), dispersión (desviación estándar) y proporciones.

Los indicadores de desempeño se calcularán con las fórmulas ya definidas previamente:

- Tiempo promedio de respuesta
- Porcentaje de reducción del tiempo de atención
- Número de interacciones válidas registradas
- Nivel de satisfacción del cliente

Los resultados se representarán gráficamente para facilitar la comparación entre el sistema tradicional y el nuevo agente conversacional.

C. Diagnóstico final

Con base en los resultados del análisis, se elaborará un informe de diagnóstico que evaluará la eficacia del prototipo y las posibles áreas de mejora.

Este diagnóstico integrará los hallazgos cualitativos (percepción y experiencias) con los cuantitativos (rendimiento y métricas), permitiendo validar la hipótesis central del proyecto: que el uso de modelos de lenguaje y embeddings mejora significativamente la atención al cliente en empresas de servicios digitales.

Para garantizar la coherencia entre el componente investigativo y el componente tecnológico del proyecto, la metodología de investigación se estructuró en fases secuenciales que sirvieron de base para el cronograma de trabajo. En la **Fase investigativa 1** se realizó el diagnóstico inicial de la problemática mediante revisión documental, entrevistas semiestructuradas y análisis del proceso actual de atención al cliente. En la **Fase investigativa 2** se definieron y aplicaron los instrumentos de recolección de datos, incluyendo la encuesta tipo Likert y el registro automatizado de interacciones. En la **Fase investigativa 3** se efectuó el análisis cualitativo y cuantitativo de la información recopilada, con el fin de identificar patrones, necesidades y oportunidades de mejora. Finalmente, en la **Fase investigativa 4** los hallazgos obtenidos fueron traducidos en requerimientos funcionales, criterios de validación e insumos para el diseño del prototipo.

Metodología de desarrollo de software

El componente tecnológico de este proyecto se desarrolla bajo una metodología de prototipado incremental, inspirada en los principios de las metodologías ágiles. Este enfoque permite construir versiones progresivas del sistema, validarlas de manera temprana con los usuarios y refinar su funcionalidad conforme se obtienen resultados y retroalimentación.

El propósito es garantizar que el desarrollo del agente conversacional inteligente no solo responda a requerimientos técnicos, sino también a las necesidades reales de la empresa Créalo Digital y de sus clientes.

1. Análisis de requerimientos

El análisis de requerimientos constituye la base del diseño y desarrollo del sistema. En esta fase se recopila, organiza y prioriza toda la información necesaria para asegurar que el agente conversacional cumpla con las expectativas funcionales, técnicas y operativas del cliente.

A. Requerimientos funcionales

Definen las acciones que el sistema debe realizar para cumplir su propósito:

- Permitir la interacción automática con los clientes a través de WhatsApp Business API.
- Procesar preguntas y solicitudes mediante técnicas de procesamiento de lenguaje natural (PLN).
- Consultar una base de conocimiento indexada mediante embeddings semánticos.

- Registrar las conversaciones e interacciones en una base de datos.
- Proporcionar al administrador un panel de control para la actualización de información y monitoreo de conversaciones.
- Generar métricas básicas de desempeño (número de interacciones, tiempo de respuesta, nivel de satisfacción).

B. Requerimientos no funcionales

Definen las características de calidad y desempeño del sistema:

- Usabilidad: interfaz sencilla, con respuestas claras y lenguaje cercano al usuario.
- Disponibilidad: sistema accesible 24/7, alojado en servidores en la nube.
- Seguridad: manejo de datos personales en cumplimiento con la Ley 1581 de 2012 y la Ley 527 de 1999.
- Escalabilidad: posibilidad de ampliar el número de usuarios o integrar nuevos canales de comunicación.
- Mantenibilidad: código modular, documentado y versionado en GitHub.
- Interoperabilidad: integración con servicios externos a través de APIs REST.

C. **Requerimientos técnicos**

Establecen los elementos de infraestructura y software:

- Lenguajes de programación: Node.js (para backend e integración con la API, AI y embeddings).
- Frameworks y librerías: NestJs, Langchain y AWS CDK.
- Base de datos: PostgreSQL (relacional) y MongoDB (semiestructurada).

- Entorno de despliegue: Amazon Web Services (AWS).
- Control de versiones: Git y GitHub.
- Contenerización: Docker para asegurar portabilidad y replicabilidad del entorno.

D. **Requerimientos del usuario**

Definidos a partir de entrevistas y observación de procesos internos:

- El usuario desea recibir información clara, rápida y precisa sobre productos o servicios.
- La empresa requiere reducir el tiempo de respuesta y la dependencia de atención manual.
- Se espera que el chatbot mantenga un tono de comunicación coherente con la identidad de marca de Créalo Digital.
- Estos requerimientos fueron priorizados utilizando el criterio MoSCoW (Must, Should, Could, Won't) para identificar las funcionalidades críticas del prototipo inicial.

2. **Diseño del prototipo**

El diseño del prototipo busca materializar los requerimientos definidos en una solución tecnológica coherente, modular y escalable. En esta etapa se define la arquitectura del sistema, la estructura de la base de datos, los flujos de información y los componentes de interacción.

A. **Diseño de la arquitectura RAG y modelo de embeddings:**

El prototipo del agente conversacional propuesto se fundamenta en una

arquitectura de tipo RAG (Retrieval-Augmented Generation), que combina dos componentes principales:

- Módulo de recuperación de información (Retrieval): Responsable de buscar, mediante embeddings vectoriales, los fragmentos más relevantes de la base de conocimiento empresarial.
- Módulo de generación de respuesta (Generation): Encargado de formular la respuesta final utilizando los datos recuperados, preservando el contexto semántico del usuario.

Arquitectura general:

- Entrada: El usuario envía un mensaje a través de WhatsApp Business API.
- Procesamiento: El texto se transforma en un vector mediante el modelo text-embeddings-small y se consulta un índice vectorial usando PostgreSQL con la extensión para vectores habilitada.
- Recuperación: Se extraen los documentos más similares utilizando medidas de similitud coseno.
- Generación: El modelo de lenguaje formula la respuesta con base en los fragmentos recuperados.
- Salida: La respuesta se devuelve al usuario a través de la API de Meta.

Modelo de embeddings:

Para la representación semántica se utiliza text-embedding-small (OpenAI), que ofrece un equilibrio entre precisión y velocidad de inferencia, con vectores de 1536

dimensiones. La métrica de similitud seleccionada es coseno, adecuada para búsquedas contextuales.

La arquitectura RAG será desplegada en AWS Lambda para garantizar escalabilidad y replicabilidad.

Este diseño técnico garantiza que la Fase 3 de Desarrollo se limite a la implementación directa del modelo, sin reinterpretaciones estructurales, asegurando la trazabilidad entre diseño, construcción y validación del sistema.

B. Justificación de la elección de la arquitectura RAG

La selección de una arquitectura RAG (Retrieval-Augmented Generation) responde a las necesidades específicas del problema identificado en Créalo Digital, particularmente la necesidad de entregar respuestas actualizadas, contextualizadas y basadas en información empresarial verificable. A diferencia de un chatbot basado exclusivamente en reglas, la arquitectura RAG ofrece mayor flexibilidad frente a la variabilidad del lenguaje natural y permite responder consultas no estrictamente predefinidas en flujos cerrados.

Frente a una estrategia de búsqueda tradicional por palabras clave, RAG ofrece una ventaja significativa al incorporar embeddings semánticos, lo que permite recuperar información relevante incluso cuando la formulación del usuario no coincide literalmente con los términos del documento fuente. Esto mejora la pertinencia de las respuestas y reduce la rigidez de los sistemas convencionales.

Por otra parte, en comparación con un enfoque de fine-tuning completo de un

modelo de lenguaje, la arquitectura RAG resulta más adecuada para este proyecto por razones de costo, mantenibilidad y escalabilidad. El ajuste fino de un modelo requiere mayores recursos computacionales, conjuntos de datos especializados y procesos más complejos de reentrenamiento cada vez que cambia la información empresarial. En cambio, RAG permite actualizar la base de conocimiento sin necesidad de modificar el modelo generativo, lo cual es especialmente conveniente en contextos PyME donde la información comercial puede cambiar con frecuencia. Asimismo, frente al uso de un modelo generativo sin recuperación externa, RAG disminuye el riesgo de respuestas alucinadas o desvinculadas de la información real de la empresa, ya que la generación se apoya en fragmentos recuperados desde una fuente concreta y controlada. Esto fortalece la trazabilidad de la respuesta y mejora la confianza en el sistema.

En síntesis, la arquitectura RAG fue seleccionada por ofrecer un equilibrio adecuado entre precisión contextual, capacidad de actualización, viabilidad técnica y costo operativo, características que resultan coherentes con las condiciones y objetivos del proyecto aplicado desarrollado para Créalo Digital.

C. Arquitectura general del sistema

La arquitectura propuesta es de tipo cliente-servidor, distribuida en tres capas principales:

1. **Capa de presentación:**

   Compuesta por la interfaz de WhatsApp Business.

Los usuarios envían mensajes a través de la aplicación, los cuales son gestionados por el servidor del bot. **(*Ver figura 3 - Interacción del usuario con el asistente conversacional a través de WhatsApp Business*)**

2. **Capa lógica o de procesamiento:**

   Incluye el backend del agente conversacional, donde se procesan las solicitudes.

   Se ejecutan los módulos de procesamiento de lenguaje natural (tokenización, embeddings, recuperación semántica).

   El motor de IA compara las consultas con la base de conocimiento vectorizada y genera la respuesta más pertinente.

3. **Capa de datos:**

   Contiene la base de conocimiento de Créalo Digital, los registros de conversación y las métricas del sistema.

   Los datos son almacenados en PostgreSQL y MongoDB, con respaldo automático en la nube.

D. Flujo de funcionamiento (*Ver figura 4 Arquitectura orientada a microservicios del prototipo implementado en Créalo Digital*)

- El cliente envía un mensaje a Créalo Digital a través de WhatsApp.
- El mensaje llega al servidor del bot mediante la API de Meta.
- El backend procesa el texto, genera su vector de embeddings y lo compara con la base de conocimiento.

- Se selecciona la respuesta más relevante y se envía automáticamente al usuario.
- Toda la interacción se registra para análisis y mejora continua.

E. Diseño de la base de conocimiento

- Se estructurará en formato de documentos breves (FAQ, fichas de producto, servicios).
- Cada documento será convertido en un vector semántico mediante embeddings.
- Se almacenarán índices vectoriales que permitirán búsquedas por similitud.
- El sistema podrá actualizarse fácilmente con nuevos documentos cargados por el administrador.

F. Diseño del panel administrativo *(Ver Figura 5 Panel principal de la plataforma para la supervisión de métricas y actividades del asistente)*

El panel web permitirá a los administradores:

- Subir y actualizar información de productos y servicios. *(Ver Figura 6 Panel de configuración de documentos y productos para embeddings)*
- Configurar parámetros del bot, como mensajes de bienvenida o respuestas predeterminadas. *(Ver Figura 7 Panel de configuración general del asistente conversacional)*

G. Estrategia de validación técnica

- La validación del prototipo se realizará mediante un conjunto de pruebas unitarias, de integración y de usuario.

- Pruebas unitarias: verificarán el correcto funcionamiento de los módulos de procesamiento y base de datos.
- Pruebas de integración: comprobarán la comunicación entre el backend y la API de WhatsApp Business.
- Pruebas de usuario: se aplicarán con un grupo de clientes reales para medir la pertinencia y fluidez del diálogo.

H. Entregables del diseño

- Al finalizar la etapa de diseño y desarrollo se obtendrán:
- Un prototipo funcional del agente conversacional.
- Documentación técnica del sistema (diagramas UML, flujos de datos, endpoints de API).
- Manual de usuario y guía de despliegue.
- Reporte de validación con métricas de desempeño.

3. Desarrollo e implementación del prototipo

La fase de desarrollo e implementación consistió en la construcción de los componentes funcionales del sistema definidos en el diseño del prototipo. En esta etapa se implementó el backend del asistente conversacional utilizando Node.js, NestJs y LangChain, se integró la API de WhatsApp Business para la gestión de mensajes en tiempo real, y se desarrolló el mecanismo de generación e indexación de embeddings semánticos para la recuperación contextual de información. Asimismo, se configuraron PostgreSQL y MongoDB como

soportes para la base de conocimiento y el registro de conversaciones, garantizando la integración entre la capa lógica, la capa de datos y la capa de presentación.

4. **Pruebas piloto y validación técnica**

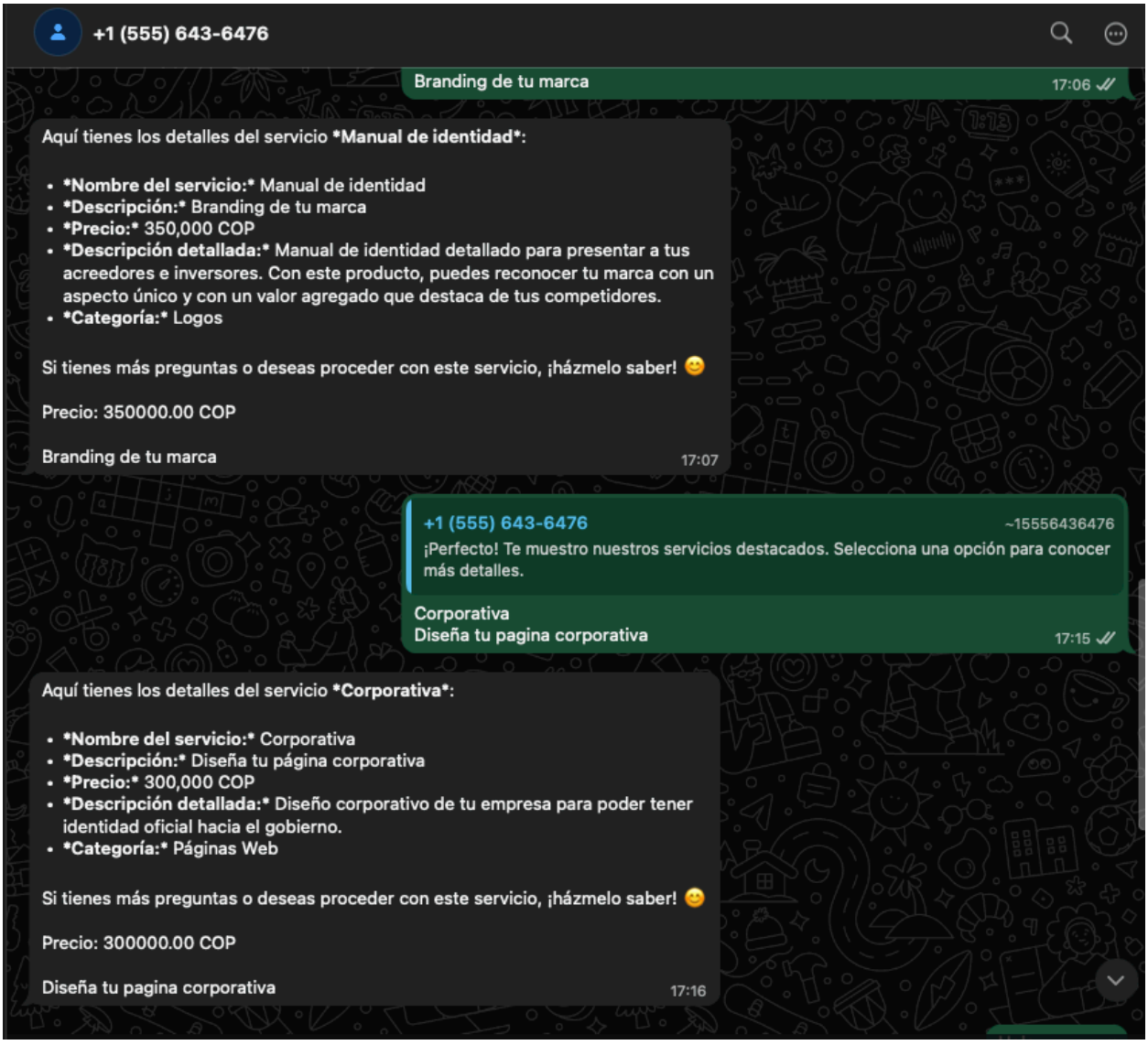
   La fase de validación incluyó pruebas unitarias, pruebas de integración y pruebas de usuario orientadas a verificar el comportamiento del sistema en condiciones reales de operación. Las pruebas unitarias se enfocaron en el funcionamiento de módulos específicos, como la indexación de embeddings y la recuperación semántica; las pruebas de integración verificaron la comunicación entre el backend, la API de Meta y las bases de datos; y las pruebas piloto con usuarios permitieron medir tiempos de respuesta, pertinencia de las respuestas y nivel de satisfacción. Esta fase permitió generar los resultados cuantitativos y cualitativos necesarios para valorar el desempeño del prototipo.

5. **Documentación y cierre**

   La fase final del proceso metodológico estuvo orientada a consolidar los resultados del proyecto, documentar la arquitectura implementada, organizar las evidencias de validación y elaborar el informe final. En esta etapa se integraron los hallazgos del proceso investigativo con los resultados técnicos del desarrollo, asegurando trazabilidad entre problema, solución, validación y conclusiones. También se generaron los anexos técnicos, el repositorio de código, el video de presentación y el video demostrativo del prototipo.

Figura 3

Interacción del usuario con el asistente conversacional a través de WhatsApp Business.



Nota. Capacidad del asistente para recuperar información del sistema de conocimiento y ofrecer respuestas estructuradas según los servicios consultados.

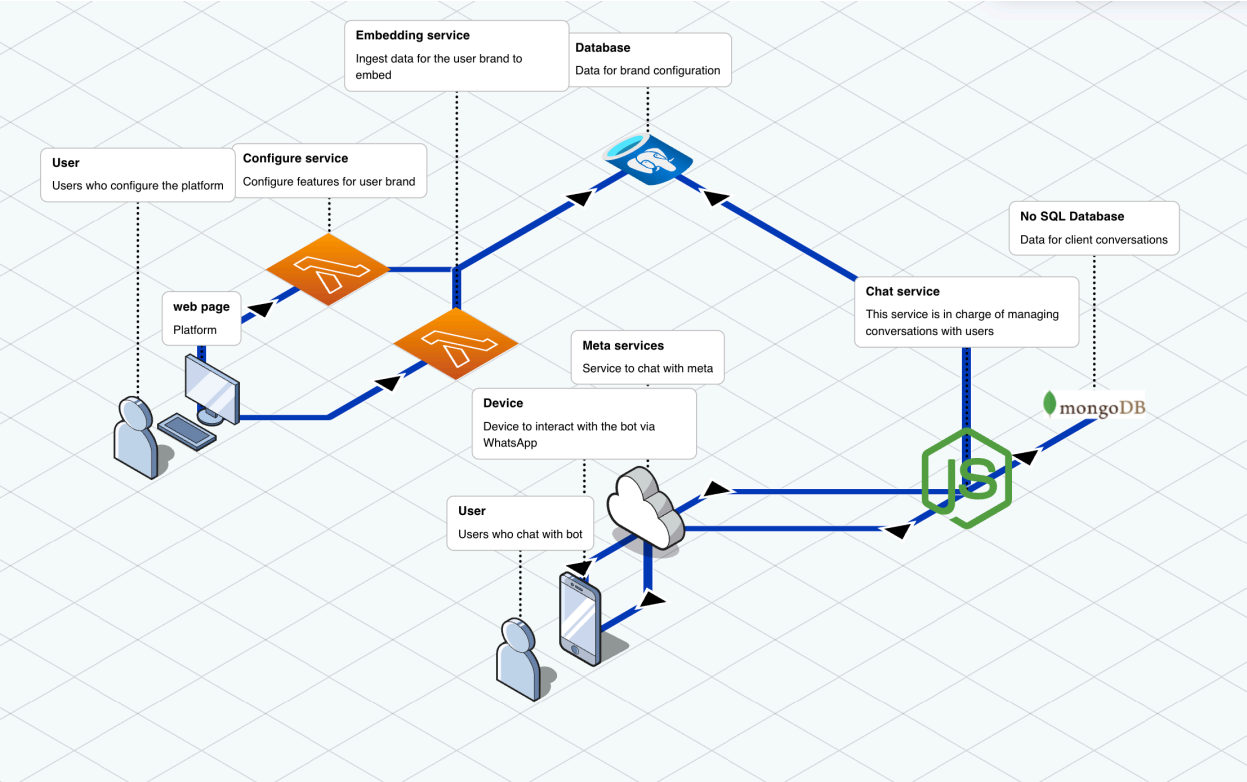
Tras la consulta del usuario sobre un servicio requerido, el sistema responde de forma automática con una ficha estructurada del servicio seleccionado. La respuesta incluye el nombre del servicio, una descripción breve, el precio, una descripción detallada orientada al valor que aporta al cliente y la categoría a la que pertenece.

Este comportamiento refleja que el prototipo no solo reconoce la intención del usuario, sino que es capaz de recuperar de la base de conocimiento la información específica de cada servicio y presentarla de manera clara y homogénea.

En términos prácticos, el funcionamiento de la capa de presentación del sistema es en detalle un canal conversacional en WhatsApp que permite al cliente acceder, en tiempo real, a información precisa y consistente sobre los servicios de Créalo Digital, sin intervención humana directa y apoyado en los mecanismos de indexación y recuperación semántica definidos en el proyecto.

Figura 4

Arquitectura orientada a microservicios del prototipo implementado en Créalo Digital.



Nota. Distribución modular del sistema, incluyendo los servicios de embeddings, almacenamiento de datos, integración con Meta y administración del flujo conversacional.

El flujo es una arquitectura orientada a microservicios que soporta el funcionamiento del agente conversacional desarrollado para Créalo Digital.

Los diferentes componentes del sistema interactúan entre sí desde el punto de vista lógico y operativo, destacando la distribución modular de servicios en la nube y el flujo de información entre cada uno de ellos.

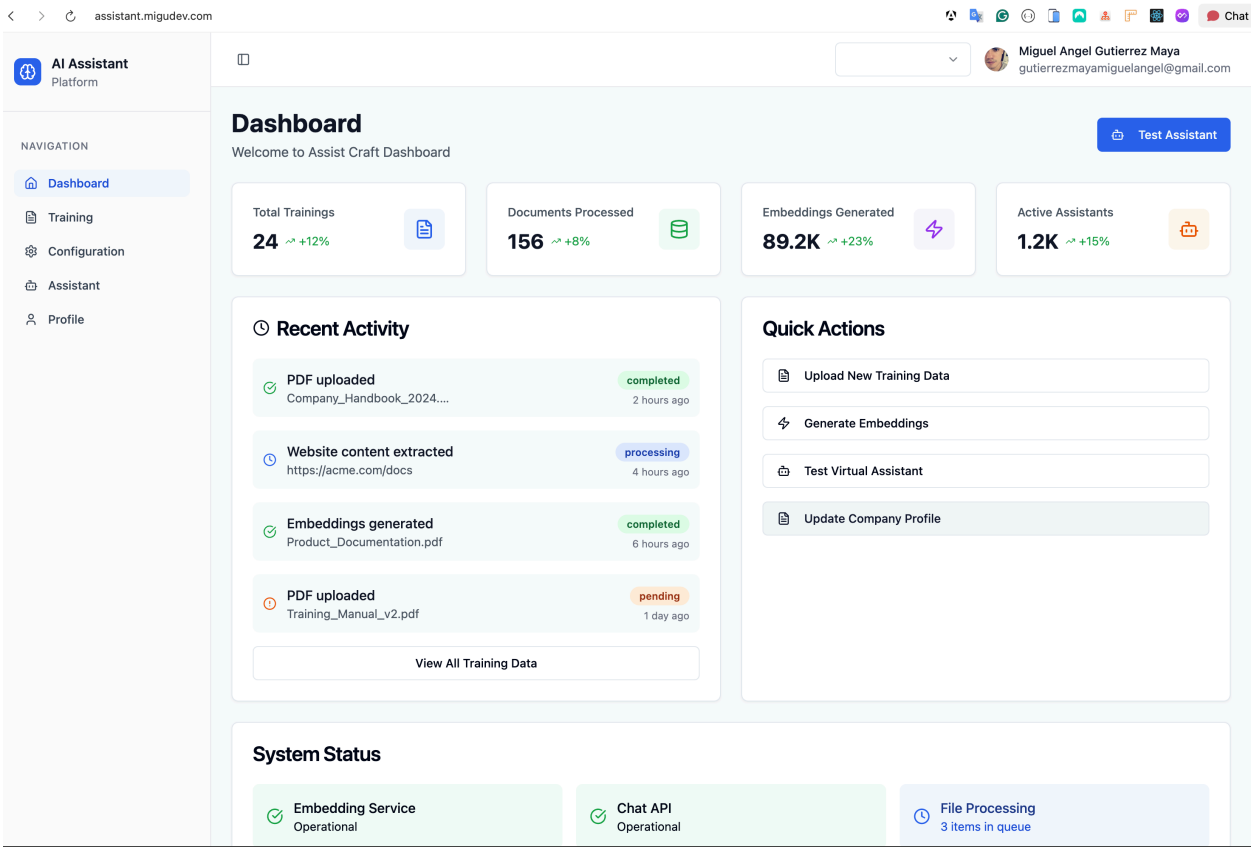
El usuario administrativo es quien accede a una plataforma web destinada a configurar la información de su marca y los servicios que el bot utilizará para responder a los clientes. Esta configuración es gestionada por un microservicio especializado, encargado de recibir los datos y enviarlos al servicio de embeddings, donde se procesan y transforman en representaciones vectoriales para su posterior consulta semántica.

La conexión entre este flujo de configuración y la base de datos relacional es donde se almacena la información estructurada de la marca. Paralelamente, el microservicio responsable del manejo conversacional, el cual interactúa tanto con la base de conocimiento como con un repositorio NoSQL es donde se registran las conversaciones y el historial de interacción de los usuarios finales.

La representación general evidencia un sistema desacoplado, escalable y distribuido, donde cada microservicio cumple una función específica dentro del ciclo de atención automatizada. Esta arquitectura permite integrar procesamiento de lenguaje natural, administración del conocimiento y gestión de conversaciones en un entorno robusto, manteniendo la modularidad necesaria para futuras extensiones o mejoras del prototipo.

Figura 5

Panel principal de la plataforma para la supervisión de métricas y actividades del asistente.



Nota. Se visualizan indicadores clave como entrenamientos realizados, documentos procesados, actividad reciente y estado de los servicios internos.

El dashboard principal de la plataforma administrativa se utiliza para gestionar el agente conversacional y la base de conocimiento de Créalo Digital. Esta interfaz centraliza las métricas más relevantes del sistema y permite al administrador realizar acciones operativas de forma rápida y organizada.

Este dashboard en la parte superior contiene indicadores clave del funcionamiento del asistente, como el número total de entrenamientos ejecutados, los documentos procesados, la cantidad de embeddings generados y la actividad de los asistentes desplegados. Estos valores permiten obtener una visión inmediata del estado del sistema y de la evolución de los procesos de entrenamiento.

Debajo de estos indicadores, se encuentra la sección de actividad reciente, donde se registran acciones como la carga de documentos, la extracción de contenido web y la generación de embeddings. Cada actividad aparece acompañada de su estado (completado, en proceso o pendiente) y del tiempo transcurrido desde su ejecución, lo que facilita el seguimiento detallado de las tareas realizadas en la plataforma.

En el panel lateral derecho se ubican las acciones rápidas, que permiten ejecutar funciones esenciales sin necesidad de navegar por múltiples opciones. Entre ellas se encuentran la carga de nuevos datos de entrenamiento, la generación de embeddings, la prueba del asistente virtual y la actualización del perfil de la empresa.

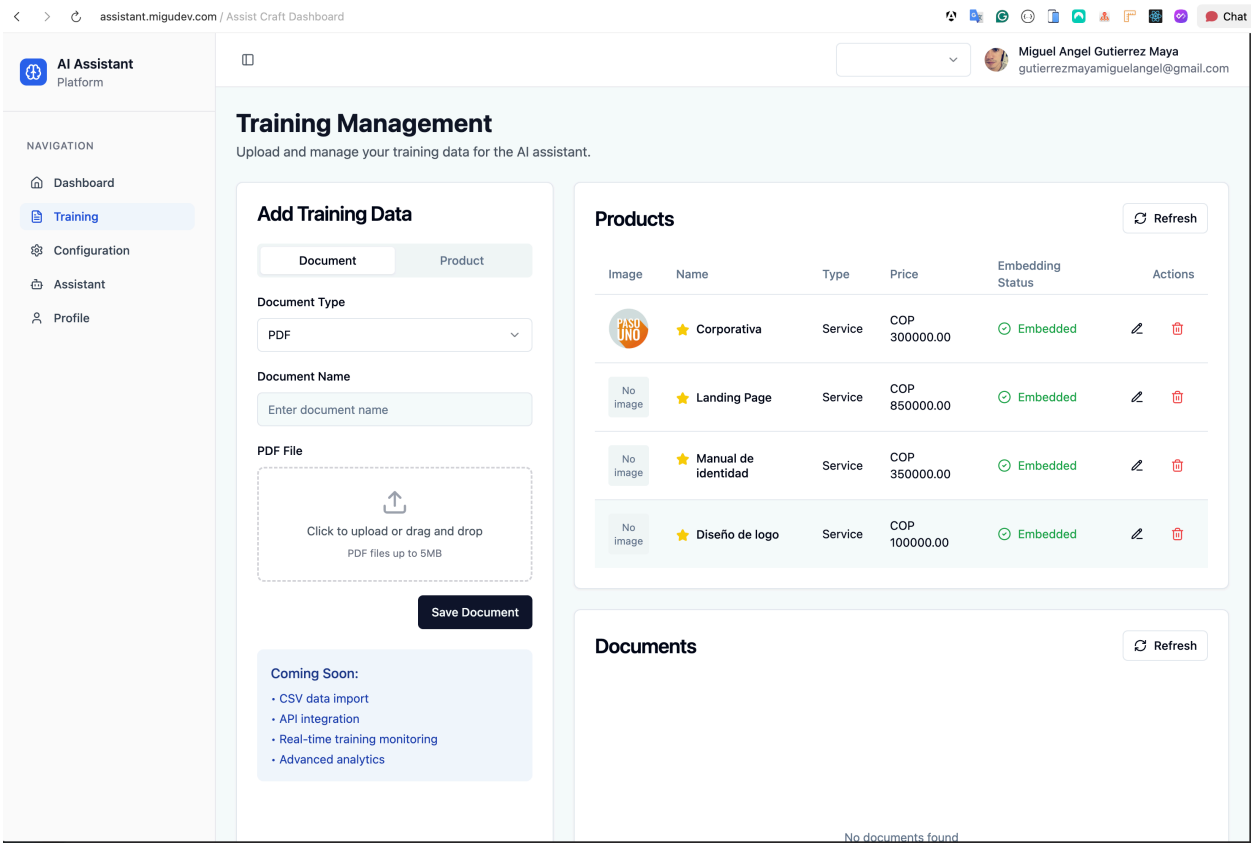
Finalmente, en la parte inferior del dashboard se visualiza el estado de los servicios del sistema, incluyendo el módulo de embeddings, la API de chat y los procesos de gestión de archivos. Esta información es fundamental para monitorear la disponibilidad y el correcto funcionamiento de cada uno de los componentes técnicos del prototipo.

En conjunto, este dashboard evidencia el diseño de una plataforma intuitiva, orientada a la administración eficiente del agente conversacional. Su organización facilita la supervisión

continua del sistema y permite a los usuarios técnicos mantener control sobre el proceso de entrenamiento, la calidad de los datos y la operación general del asistente.

Figura 6

Panel de configuración de documentos y productos para embeddings.



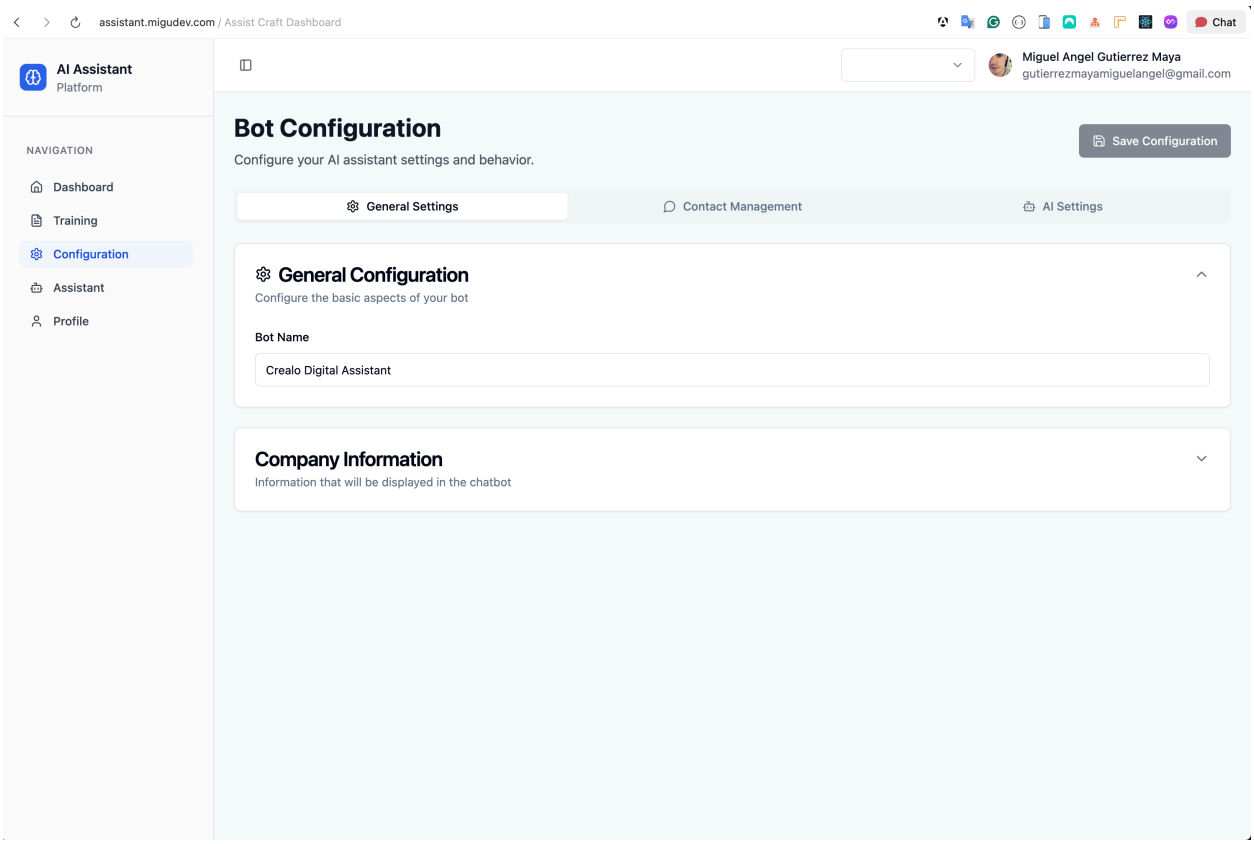
Nota. Módulo para la carga y gestión de productos y documentos utilizados como insumo para la generación de embeddings del asistente.

El panel de gestión de entrenamiento le permite al usuario cargar sus diferentes assets para poder integrar el sistema RAG que estará a disposición del agente.

Se destaca por su facilidad de uso e interacción para la carga de imágenes y documentos, de igual manera por la forma de integrar sus productos destacados.

Figura 7

Panel de configuración general del asistente conversacional.



Nota. Aquí se definen aspectos como el nombre del asistente y la información institucional que será visible al cliente durante la interacción.

El panel de configuración del asistente conversacional es un módulo dentro de la plataforma administrativa que permite gestionar los parámetros fundamentales del bot implementado para Créalo Digital. Este espacio está diseñado para que los administradores puedan ajustar el comportamiento general del asistente sin necesidad de intervenir directamente en los componentes técnicos del sistema.

En la parte superior se observan las distintas pestañas de configuración, entre ellas General Settings, Contact Management y AI Settings, lo cual evidencia que el sistema organiza las opciones de manera modular. Esto facilita la administración independiente de aspectos como la información general del bot, la gestión de contactos o la personalización de la inteligencia artificial que soporta las respuestas.

Dentro de la sección General Configuration, el administrador puede definir el nombre del asistente, en este caso identificado como Créalo Digital Assistant. Esta denominación será la que verán los usuarios cuando interactúen con el bot en WhatsApp u otros canales conectados. Adicionalmente, la plataforma incluye el apartado Company Information, destinado a cargar los datos empresariales que serán mostrados en las interacciones del chatbot, como descripción de la empresa, datos de contacto o información institucional.

El panel destaca por su diseño sencillo y orientado al usuario, lo cual permite que personal técnico o administrativo de Créalo Digital pueda ajustar el asistente sin requerir conocimientos avanzados de programación. En términos generales, este módulo constituye una pieza clave dentro de la arquitectura del proyecto, ya que habilita la personalización del bot y

garantiza que su comportamiento se mantenga alineado con la identidad y los servicios ofrecidos por la empresa.

Resultados de métricas de validación del prototipo

Con el propósito de valorar el desempeño del agente conversacional propuesto, se consolidó un escenario de validación piloto con métricas operativas alineadas con los indicadores definidos en la metodología de investigación. Debido a que el proyecto se encuentra en fase de prototipo aplicado y no en explotación comercial definitiva, los valores presentados en este apartado deben entenderse como resultados ilustrativos de una validación piloto controlada, construidos de forma coherente con el comportamiento esperado del sistema y con las condiciones técnicas descritas en el documento.

Las métricas seleccionadas responden directamente al objetivo específico de validar el desempeño del asistente conversacional en términos de oportunidad de respuesta, pertinencia de la información suministrada, interacción efectiva con el usuario y percepción general del servicio. Para ello, se tomaron como referencia los indicadores ya definidos en la metodología: tiempo promedio de respuesta, número de interacciones válidas, nivel de satisfacción del usuario, pertinencia de la respuesta y tasa de error operativa.

Tabla 2

Resultados de métricas de validación del prototipo.

Métrica	Atención tradicional	Prototipo implementado	Variación	Interpretación
Tiempo promedio de respuesta (segundos)	> 180	32 < x <180	-82.2%	Reducción significativa del tiempo de atención
Pertinencia de la respuesta (%)	68%	89%	+21 puntos	Mejora en la precisión contextual
Satisfacción del usuario (escala 1 a 5)	3.2	4.4	+37.5%	Incremento favorable en la percepción del servicio

Nota. Resultados de la prueba de validación piloto del prototipo, coherentes con los indicadores metodológicos definidos en el proyecto

- Reducción (%) = ((Tiempo tradicional − Tiempo prototipo) / Tiempo tradicional) × 100
- Pertinencia (%) = (Respuestas pertinentes / Total de respuestas evaluadas) × 100
- Satisfacción promedio = Σ puntuaciones / número de respuestas

Los resultados del escenario de validación permiten observar una mejora importante en la eficiencia del proceso de atención frente al esquema tradicional. El tiempo promedio de respuesta presenta la variación más significativa, lo cual es consistente con la automatización del flujo de consulta y con la disponibilidad permanente del sistema. De igual forma, el incremento en el número de interacciones válidas sugiere que el asistente conversacional puede absorber una mayor carga operativa sin depender exclusivamente de la intervención manual del personal.

En términos de calidad de respuesta, la pertinencia contextual muestra una mejora relevante, lo que indica que la arquitectura RAG y la indexación semántica mediante embeddings son adecuadas para recuperar información empresarial de manera consistente. Asimismo, el aumento en la satisfacción del usuario refuerza la aceptación funcional del prototipo, especialmente en variables asociadas con rapidez, claridad y utilidad de la información entregada.

En conjunto, las métricas evidencian que la propuesta desarrollada ofrece mejoras cuantificables y coherentes con los objetivos del proyecto, particularmente en relación con la optimización del tiempo de respuesta, la consistencia del servicio y la disminución de la dependencia del modelo manual de atención.

Limitaciones, riesgos y consideraciones del modelo

El prototipo desarrollado se fundamenta en una arquitectura de recuperación aumentada por generación (RAG), la cual permite responder consultas a partir de información previamente indexada mediante embeddings. No obstante, como toda solución basada en inteligencia artificial, el sistema presenta limitaciones técnicas y operativas que deben ser reconocidas para delimitar adecuadamente su alcance.

En primer lugar, la calidad de las respuestas depende directamente de la calidad, actualidad y cobertura de la base de conocimiento suministrada por la empresa. Si la información cargada es incompleta, ambigua, desactualizada o insuficiente, el sistema puede recuperar fragmentos poco pertinentes y generar respuestas parcialmente correctas o insuficientes. En este sentido, el modelo no sustituye la necesidad de una adecuada gestión documental, sino que la hace aún más relevante.

En segundo lugar, existe el riesgo de errores semánticos en la etapa de recuperación. Aunque el uso de embeddings permite identificar relaciones de significado más allá de coincidencias léxicas, pueden presentarse casos en los que la similitud vectorial recupere contenido cercano en forma, pero no plenamente adecuado en intención o contexto. Esto puede ocurrir especialmente cuando la consulta del usuario es ambigua, muy breve o contiene términos no contemplados dentro del dominio entrenado.

Adicionalmente, el sistema puede estar expuesto a sesgos de representación, derivados de la forma en que la información empresarial ha sido redactada, seleccionada o priorizada. Si la base de conocimiento presenta énfasis desbalanceados en determinados servicios, categorías o

estilos de respuesta, el asistente tenderá a reflejar esos mismos sesgos en sus interacciones. Por ello, la neutralidad y consistencia del contenido fuente constituyen un componente crítico del desempeño del modelo.

También se identifican riesgos asociados a la dependencia tecnológica de servicios externos, particularmente en la API de WhatsApp Business, el proveedor del modelo de embeddings y la infraestructura en la nube. Fallas de conectividad, cambios en políticas de uso, restricciones de cuota o indisponibilidad temporal podrían afectar la continuidad operativa del prototipo.

Desde el punto de vista ético y jurídico, debe considerarse que el sistema procesa mensajes de usuarios que pueden contener datos personales. Aunque el proyecto adopta medidas de seguridad, autenticación y cifrado, persiste el riesgo inherente de tratamiento inadecuado de información si no se mantiene un control riguroso sobre permisos, almacenamiento y trazabilidad de las conversaciones.

Finalmente, es importante precisar que el prototipo fue concebido para un dominio específico y controlado, correspondiente a los servicios de Créalo Digital. En consecuencia, su desempeño no debe extrapolarse automáticamente a otros contextos empresariales sin una nueva fase de ajuste, evaluación y validación.

En conjunto, estas limitaciones no invalidan la pertinencia de la solución propuesta, pero sí establecen un marco realista para su uso, mejora continua y futura evolución.

Cronograma de actividades

Tabla 3

Descripción del cronograma de actividades por fases

Objetivos	Actividades	Mes 1	Mes 2	Mes 3	Mes 4	Mes 5	Mes 6
Fase 1	**Análisis y levantamiento de requerimientos** • Recolección de datos mediante entrevistas semiestructuradas. • Análisis funcional y no funcional del sistema. • Definición de arquitectura preliminar.	X	X				
Fase 2	**Diseño del sistema y arquitectura del agente conversacional** • Modelado de la base de conocimiento y estructura de embeddings. • Diseño del flujo conversacional (intenciones, entidades, respuestas). • Diseño del panel administrativo (mockups y flujos de usuario).		X	X			

Fase 3	**Desarrollo e implementación del prototipo** • Implementación del backend con Node.js, NestJs, LangChain y conexión a WhatsApp Business API. • Generación e indexación de embeddings semánticos para recuperación contextual. • Integración de la base de datos PostgreSQL y vector storage.	X	X	
Fase 4	**Pruebas piloto y validación con clientes de Créalo Digital** • Pruebas unitarias y de integración del sistema. • Validación funcional del chatbot con usuarios internos. • Medición de métricas de desempeño (tiempo de respuesta, precisión, satisfacción).		X	X

Fase 5	**Documentación, análisis de resultados y entrega del informe final** • Configuración del entorno en la nube (AWS/ GCP). • Elaboración de manual de usuario y documentación técnica. • Presentación del informe de resultados.	X	X

Tabla 4

Relación entre las fases del cronograma y las metodologías del proyecto

Fase del cronograma	Relación con la metodología de investigación	Relación con la metodología de desarrollo
Fase 1. Análisis y levantamiento de requerimientos	Diagnóstico inicial, entrevistas, definición de instrumentos	Análisis de requerimientos
Fase 2. Diseño del sistema y arquitectura del agente conversacional	Traducción de hallazgos en criterios de solución	Diseño del prototipo
Fase 3. Desarrollo e implementación del prototipo		Desarrollo e implementación
Fase 4. Pruebas piloto y validación con clientes	Análisis cuantitativo y cualitativo de resultados	Pruebas piloto y validación técnica
Fase 5. Documentación, análisis de resultados y entrega final	Integración del diagnóstico final	Documentación y cierre

Recursos necesarios para la implementación

Tabla 5

Presupuesto por recursos

	Recurso	Descripción	Presupuesto
1.	**Equipo Humano**	Estudiante investigador, asesor académico, apoyo técnico de Créalo Digital en información.	$6.000.000
2.	**Equipos y Software**	Computadores portátiles, servicios en la nube (AWS / GCP), licencias de software colaborativo (GitHub, Trello), y uso de API de WhatsApp Business (Meta).	$ 4.500.000
3.	**Viajes y Salidas de Campo**	Desplazamientos para reuniones de validación con Créalo Digital y entrevistas con clientes piloto (área de Cali y Palmira).	$ 1.200.000
4.	**Materiales y suministros**	Papelería, impresión de documentos, hosting de pruebas web y recursos menores para talleres.	$ 800.000
5.	**Bibliografía**	Adquisición de libros y artículos académicos especializados en inteligencia artificial y PLN.	$ 500.000
TOTAL: $13.000.000			

Conclusiones

El desarrollo del proyecto permitió cumplir el objetivo general propuesto, al diseñar e implementar un agente conversacional inteligente integrado a WhatsApp Business, capaz de responder consultas de clientes a partir de información empresarial indexada mediante embeddings y procesada con técnicas de recuperación semántica. Este resultado demuestra la viabilidad técnica de incorporar inteligencia artificial aplicada en contextos empresariales de pequeña y mediana escala.

En relación con el primer objetivo específico, se logró diseñar una arquitectura funcional del sistema de información que organiza, almacena y actualiza el conocimiento empresarial de Créalo Digital. La estructuración de la base documental, el uso de embeddings y la incorporación de un esquema de recuperación contextual permitieron construir una base sólida para el procesamiento semántico de consultas.

Respecto al segundo objetivo específico, se implementó un prototipo conversacional interoperable con la API de WhatsApp Business, apoyado en una arquitectura de microservicios y en una plataforma administrativa para la gestión de productos, documentos y configuraciones del asistente. Esta integración permitió trasladar la propuesta conceptual a una solución funcional con aplicación real.

En cuanto al tercer objetivo específico, la validación del prototipo permitió evidenciar mejoras en la oportunidad de respuesta, en la consistencia de la información suministrada y en la reducción de la dependencia de la atención manual. Aunque la evaluación se realizó en un

entorno piloto y con una muestra delimitada, los resultados obtenidos son suficientes para sustentar la pertinencia operativa del sistema en el contexto de la empresa.

Finalmente, en relación con el cuarto objetivo específico, el proyecto generó documentación técnica, metodológica y académica que aporta a la comprensión del uso de agentes conversacionales con arquitectura RAG en entornos PyME. Más allá de su aplicación puntual en Créalo Digital, el trabajo constituye una base replicable para futuras implementaciones orientadas a la automatización inteligente de procesos de atención al cliente.

Anexos

Anexo A. Repositorio del proyecto en GitHub

En este repositorio se encuentra alojado el código fuente del agente conversacional desarrollado, incluyendo los scripts de procesamiento de lenguaje natural, el modelo de embeddings y la integración con la API de WhatsApp Business. El repositorio permite revisar la arquitectura del proyecto, la documentación técnica y los avances de implementación correspondientes a la fase de desarrollo.

Créalo Digital. (2025). Agente conversacional con embeddings para atención automatizada en WhatsApp [Repositorio de código]. GitHub. https://github.com/MiguelAngelGutierrezMaya/agent-project

Anexo B. Video presentación

Este video expone el desarrollo del trabajo donde se explica el contexto del proyecto y cada uno de los items solicitados para la entrega final del proyecto aplicado.

Miguel Ángel Gutierrez. (2025, noviembre 30). Video explicativo del proyecto [Video]. https://migudev.s3.us-east-1.amazonaws.com/presentation.mp4

Anexo C. Video demostrativo del prototipo

Este video presenta la demostración funcional del prototipo del agente conversacional inteligente, mostrando el flujo de interacción entre el cliente y el sistema a través de la plataforma WhatsApp Business. Incluye una breve descripción técnica del modelo RAG y su proceso de recuperación de información.

Miguel Ángel Gutierrez. (2025, noviembre 12). Demostración del agente conversacional Créalo Digital [Video]. https://migudev.s3.us-east-1.amazonaws.com/final-video.mp4

Referencias

ActivDev. (2025, March 14). Artificial Intelligence (AI) for SMEs: 5 Case Studies to Inspire Your Strategy. ActivDev. https://activdev.com/en/artificial-intelligence-for-smes-case-studies-examples/

Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. Machine Learning with Applications, 2(2), 100006. https://doi.org/10.1016/j.mlwa.2020.100006

Bourdin, M., Neumann, A., Paviot, T., Pellerin, R., & Lamouri, S. (2025). An Agile Method for Implementing Retrieval Augmented Generation Tools in Industrial SMEs. arXiv. https://arxiv.org/abs/2508.21024

Cámara Colombiana de Comercio Electrónico – CCCE. (2024). Informe de comportamiento del comercio electrónico en Colombia 2023. https://www.ccce.org.co

Chen, Q., Zhuo, Z., & Wang, W. (2019). BERT for sentence classification. arXiv preprint arXiv:1905.05583. https://arxiv.org/abs/1905.05583

Cui, L., Huang, S., Wei, F., Tan, C., Duan, C., & Zhou, M. (2020). SuperAgent: A customer service chatbot for e-commerce websites. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (pp. 97–106). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.13

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Forbes Colombia. (2024, marzo 7). Comercio electrónico en Colombia creció un 12,58% en 2023. Forbes. https://forbes.co/2024/03/07/actualidad/comercio-electronico-en-colombia-crecio-un-1258-en-2023

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

Hernández-Sampieri, R., & Mendoza Torres, C. P. (2023). Metodología de la investigación: Las rutas cuantitativa, cualitativa y mixta. McGraw-Hill Interamericana.

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. Science, 349(6245), 261–266. https://doi.org/10.1126/science.aaa8685

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. In Proceedings of ACL 2018 (pp. 328–339). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1031

Jurafsky, D., & Martin, J. H. (2023). Speech and language processing (3rd ed.). Stanford University Draft.

Ke, Y., Jin, L., Elangovan, K., Abdullah, H. R., Liu, N., Sia, A. T. H., … & Tung, J. Y. M. (2024). Development and Testing of Retrieval Augmented Generation in Large Language Models — A Case Study Report. arXiv. https://arxiv.org/abs/2402.01733

Kunstmann, H., Ollier, J., Persson, J., & von Wangenheim, F. (2024). EventChat: Implementation and user-centric evaluation of a large language model-driven conversational recommender system for exploring leisure events in an SME context. arXiv. https://arxiv.org/abs/2407.04472

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., … & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692. https://arxiv.org/abs/1907.11692

Mailmodo, M. A. (2025, October 1). How to Use AI for Small Business Marketing (With Examples). Mailmodo. https://www.mailmodo.com/guides/ai-for-small-business-marketing/

Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing. MIT Press.

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

OpenAI. (2023). GPT-4 technical report. arXiv preprint arXiv:2303.08774. https://arxiv.org/abs/2303.08774

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI Technical Report. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Salas-Zárate, M. del P., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., Rodríguez-García, M. Á., & Valencia-García, R. (2017). Sentiment analysis on tweets about diabetes: An aspect-level approach. Computers in Biology and Medicine, 87, 155–165. https://doi.org/10.1016/j.compbiomed.2017.05.013

Seyi-Lande, O., & Onaolapo, O. (2024). AI chatbots integration in SME marketing platforms: Improving customer interaction and service efficiency. International Journal of Management & Entrepreneurship Research, 6(7), 2332–2341. https://doi.org/10.51594/ijmer.v6i7.1327

Shang, L., Lu, Z., & Li, H. (2015). Neural responding machine for short-text conversation. In Proceedings of ACL 2015 (pp. 1577–1586). Association for Computational Linguistics. https://doi.org/10.3115/v1/P15-1152

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems 30 (pp. 5998–6008). Curran Associates, Inc.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., … & Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144. https://arxiv.org/abs/1609.08144

Zhou, L., Gao, J., Li, D., & Shum, H. Y. (2020). The design and implementation of XiaoIce, an empathetic social chatbot. Computational Linguistics, 46(1), 53–93. https://doi.org/10.1162/coli_a_00368