



Desarrollo de Asistente Virtual con Procesamiento de Lenguaje
Natural para Servicios al Cliente en Bancos Argentinos

Profesora de Trabajo Final: Dra. Marcela Samela

Alumno: Carlos Manuel Lopez

Trabajo Final de Carrera

Lic. En Gestión de Tecnología Informática

Diciembre, 2025

Resumen

Los asistentes virtuales inteligentes se han convertido en una herramienta clave para automatizar y optimizar procesos de diversas industrias. El acceso global a tecnologías de procesamiento del lenguaje natural (PLN) ha permitido superar las limitaciones de los chatbots tradicionales, los cuales solían generar resistencia por parte de los usuarios. Actualmente, los denominados agentes virtuales inteligentes ofrecen una experiencia más cercana al razonamiento humano, brindando respuestas más precisas y una interacción más fluida con los clientes.

En el presente trabajo, se propone una intervención profesional orientada al análisis y mejora de las capacidades de atención al cliente en el sector bancario argentino. Para ello, se estudiará el estado actual de implementación de soluciones de Chatbots e Inteligencia Artificial (IA) en dicho sector, identificando oportunidades de mejora mediante el uso de la plataforma Salesforce y sus herramientas de PLN.

La propuesta contempla el diseño e implementación de un agente virtual inteligente, detallando los casos de uso sugeridos, la arquitectura del desarrollo, los recursos requeridos, los tiempos estimados de implementación, las tareas involucradas y los riesgos asociados al despliegue del sistema. El plan de implementación ofrecerá una solución que reduzca la carga de consultas que deben ser atendidas por personal humano, incrementando la eficiencia operativa y mejorando la experiencia del cliente.

Palabras clave: *agentes de software, inteligencia artificial generativa, procesamiento del lenguaje natural, servicios de atención al cliente, servicios financieros*

Abstract

Intelligent virtual assistants have become a key tool for automating and optimizing processes across various industries. The global accessibility of natural language processing (NLP) technologies has enabled businesses to overcome the limitations of traditional chatbots, which often generated user resistance. Today, so-called virtual agents offer a more human-like reasoning experience, providing more accurate responses and smoother customer interactions.

This professional intervention project aims to analyze and enhance customer service capabilities within the Argentine banking sector. It will assess the current implementation status of artificial intelligence (AI) solutions in the industry and identify opportunities for improvement using Salesforce and its NLP tools.

The proposal outlines the design and implementation of an intelligent virtual agent, including recommended use cases, development architecture, required resources, estimated implementation timeframes, associated tasks, and potential risks. The main objective is to offer a solution that reduces the volume of inquiries requiring human attention, thereby improving operational efficiency and enhancing customer experience.

Keywords: *customer services, financial services, generative artificial intelligence, natural language processing, software agents*

Dedicatoria

A mi esposa, y a mi familia, quienes me acompañaron con paciencia a lo largo de mi proceso de formación y durante la realización de este trabajo.

Reconocimientos

A la Dra. Marcela Samela, por su acompañamiento y compromiso como profesora del Trabajo Final de Carrera.

A la Universidad Abierta Interamericana, por brindarme la oportunidad de continuar mi desarrollo académico y profesional.

A mis compañeros y profesores, por haber contribuido a mi crecimiento con su apoyo durante toda la carrera.

Índice General

Resumen.....	2
Abstract	3
Dedicatoria	4
Reconocimientos	5
Índice General	6
Índice de Figuras	10
Índice de Tablas	12
Estructura General del Trabajo Final	13
Capítulos	13
Anexos	13
Acrónimos	13
Referencias.....	13
Capítulo 1 – Introducción	14
1.1 Descripción del problema	14
1.2 Justificación de la propuesta	14
1.3 Marco Institucional	15
1.4 Objetivos del Trabajo Final	16
1.4.1 Objetivo General.....	16
1.4.2 Objetivos Específicos.....	16
Capítulo 2 – Estado del Arte	17
2.1 Introducción	17
2.2 Situación Actual en Bancos de Argentina	19
2.3 Implementación de LLM en Servicios Financieros	23
2.3.1 Casos Internacionales del Dominio Financiero	23
2.3.2 Inteligencia Artificial en Atención al Cliente	29
2.4 Salesforce.....	31

2.4.1 Evolución de la Plataforma hacia Agentforce	31
2.4.2 Agentforce para Servicios Financieros	33
2.4.3 Acceso a Datos Centralizados.....	34
Capítulo 3 – Marco Teórico.....	36
3.1 Inteligencia Artificial.....	36
3.1.1 Introducción	36
3.1.2 Fundamentos de la Inteligencia Artificial.....	37
3.1.3 Aprendizaje automático	39
3.1.4 Funcionamiento de la IA Generativa	41
3.1.5 Arquitecturas de Modelos Generativos.....	44
3.1.6 Agentes Autónomos y Arquitecturas Híbridas en IA	45
3.2 Procesamiento de Lenguaje Natural	45
3.2.1 Introducción	45
3.2.2 Técnicas y Enfoques Principales en PLN	48
3.2.3 Rol del PLN en Asistentes Virtuales	49
3.3 Modelado de Lenguaje a Gran Escala	50
3.3.1 Introducción	50
3.3.2 Proceso de Preentrenamiento y Ajuste Fino	51
3.3.3 Ingeniería de Prompts y Optimización de Resultados	53
3.3.4 Evaluación de Desempeño en LLM: Métricas.....	54
3.4 Centros de Atención al Cliente	56
3.4.1 Inteligencia Artificial Multicanal.....	56
3.4.2 Resolución de Problemáticas Comunes	57
3.5 Gestión de Datos Personales en Banca y Gobernanza de IA.....	58
3.5.1 Ética Aplicada a la IA	58
3.5.2 Tratamiento de Datos Personales en Bancos de la Argentina.....	59
3.5.3 Marcos Regulatorios y Estándares Aplicables a IA en Banca	59

3.5.4 Seguridad, Sesgos y Mitigación de Riesgos en LLM	60
Capítulo 4 – Propuesta de Intervención	62
4.1 Descripción	62
4.2 Metodología	63
4.3 Plan de implementación del asistente virtual inteligente	64
4.3.1 Alcance Funcional	64
4.3.2 Etapas de Implementación	69
4.3.3 Arquitectura Técnica y Componentes	72
4.3.4 Integración con procesos bancarios	78
4.4 Recursos	80
4.4.1 Recursos Humanos	80
4.4.2 Recursos Tecnológicos	82
4.5 Costos del Proyecto	84
4.5.1 Costos del Equipo de Trabajo	84
4.5.2 Costos Técnicos	85
4.6 Cronograma de tareas	90
4.7 Métricas de Desempeño y Mejora Continua	92
4.8 Factores condicionantes	94
4.8.1 Factores internos	94
4.8.2 Factores externos	95
Capítulo 5 – Prototipo de Asistente Virtual	96
5.1 Introducción	96
5.2 Configuración del Entorno de Prueba	97
5.3 Desarrollo Inicial del Agente	97
5.3.1 Configuración Básica	97
5.3.1 Configuración del Asistente Virtual	98
5.3.2 Configuración de Acciones Personalizadas	100

5.4 Resultados de las Pruebas	111
5.5 Conclusiones del Prototipo	115
Conclusiones	117
Trabajos Futuros	119
Anexo	120
Acrónimos	124
Referencias.....	125

Índice de Figuras

Figura 1 Las cuatro generaciones del modelo de lenguaje desde la perspectiva de la capacidad de resolución de tareas	17
Figura 2 Distribución de tareas en BizFinBench según cinco dimensiones clave.....	24
Figura 3 Arquitectura del análisis del Agente financiero	26
Figura 4 Arquitectura del agente financiero	28
Figura 5 Gasto en inteligencia artificial (IA) en la industria bancaria a nivel mundial.....	30
Figura 6 Cuadrante Mágico para Plataformas de Automatización de la Fuerza de Ventas...32	
Figura 7 Representación vectorial de relaciones lingüísticas	43
Figura 8 Arquitectura de un transformador	47
Figura 9 Modelo predictivo con tokens	52
Figura 10 Capacidades de asistentes virtuales mejoradas por LLM.....	58
Figura 11 Desarrollo responsable en el ciclo de vida de sistemas con IA.....	61
Figura 12 Diagrama BPMN para Consultas Básicas	66
Figura 13 Diagrama BPMN para Gestión de Reclamos	67
Figura 14 Diagrama BPMN para Seguridad Digital.....	68
Figura 15 Diagrama BPMN para Revisión de Transacciones Financieras.....	69
Figura 16 Etapas de Implementación del Asistente Virtual	70
Figura 17 Motor de Razonamiento Atlas.....	74
Figura 18 Funcionamiento de la Capa de Confianza Einstein.....	76
Figura 19 Arquitectura de Componentes de la Solución	78
Figura 20 Presupuesto para la contratación de Data Cloud (expresado en dólares).....	87
Figura 21 Licenciamiento Recomendado de Agentforce	88
Figura 22 Computadora Recomendada para el Proceso de Migración de Datos.....	89
Figura 23 Configuración Inicial en Entorno de Desarrollo	99
Figura 24 Prompt para Validar Cliente al Iniciar una Conversación.....	101
Figura 25 Endpoint Configurado para Recibir Reclamos desde Salesforce	102
Figura 26 Rendimiento y Costo para GPT 4.1 Mini.....	110
Figura 27 Vista Previa de Respuesta Inicial al Cliente.....	111
Figura 28 Registro de Reclamo Automatizado.....	112
Figura 29 Visualización del Reclamo en la Base de Datos	113
Figura 30 Conversación para Solicitar un Reclamo	114
Figura 31 Nuevo Reclamo generado con Asistente Virtual	114

Figura 32	Recepción del Mensaje en Sistema Externo Simulado	115
Figura 33	Conversación con chat online del Banco Provincia	120
Figura 34	Conversación de chat del Banco Comafi	120
Figura 35	Conversación con El Búho, chat online del Banco Hipotecario	121
Figura 36	Conversación con chatbot del Banco Ciudad.....	121
Figura 37	Conversación con Banco Galicia	122
Figura 38	Conversación con BancoChat del Banco Macro	122
Figura 39	Conversación con chat del Banco Supervielle	123

Índice de Tablas

Tabla 1	Chatbots de atención al cliente en bancos argentinos y sus canales disponibles.....	21
Tabla 2	Características de los tipos de aprendizaje automático	40
Tabla 3	Costos de Perfiles Asignados al Proyecto	84
Tabla 4	Simulación de Parámetros Requeridos para Data Cloud	85
Tabla 5	Costo Total del Proyecto.....	90
Tabla 6	Etapas del Proyecto en Semanas	91
Tabla 7	Actividades de la Etapa 1 del Proyecto.....	92
Tabla 8	Actividades de la Etapa 2 del Proyecto.....	92

Estructura General del Trabajo Final

Capítulos

En el capítulo 1 se presenta la introducción, la relevancia profesional de la propuesta y los objetivos del trabajo, como así también se expone el contexto del sector bancario argentino en materia de transformación digital y atención al cliente.

En el capítulo 2 se comparte una revisión de antecedentes recientes sobre la aplicación de inteligencia artificial en el ámbito financiero y en procesos de atención al cliente.

En el capítulo 3 se incluye un análisis conceptual sobre inteligencia artificial generativa, modelos de lenguaje de gran escala (LLM), procesamiento de lenguaje natural (PLN), ingeniería de *prompts* (instrucciones) y los centros de atención al cliente.

El capítulo 4 contiene el desarrollo de la propuesta y el plan de implementación para la creación de un asistente virtual inteligente basado en la plataforma Salesforce. También se presenta la arquitectura tecnológica, los flujos de negocio modelados en notación BPMN, la integración de componentes tecnológicos y la estimación de costos y recursos humanos.

El capítulo 5 presenta el diseño y desarrollo de un prototipo funcional del asistente virtual propuesto. Se describe la configuración del entorno Salesforce *Developer Edition* (edición para desarrolladores), la interacción con datos y archivos y la creación de flujos conversacionales, junto con las pruebas realizadas para validar su funcionamiento.

Anexos

La sección de anexos contiene acerca de la observación de los asistentes virtuales disponibles en bancos argentinos como parte del análisis del contexto actual.

Acrónimos

Apartado dedicado a los acrónimos utilizados en el trabajo.

Referencias

Se incluye la bibliografía utilizada para la realización del trabajo.

Capítulo 1 – Introducción

1.1 Descripción del problema

En el contexto actual de transformación digital, el sector bancario argentino se encuentra frente a un desafío permanente: cómo mejorar la experiencia del cliente al mismo tiempo que se optimizan los canales de atención y se reduce la carga operativa en la gestión diaria de consultas y reclamos. Los clientes ya no aceptan demoras ni respuestas genéricas; esperan ser atendidos en el momento que se comunican y con soluciones útiles que puedan aplicarse a diferentes canales de atención con el mismo tiempo de espera y solidez de respuesta.

Muchas entidades bancarias argentinas han incorporado asistentes virtuales para responder a esta demanda, pero en la práctica la mayoría funciona sobre estructuras rígidas. Estos sistemas dependen de árboles de decisión predefinidos que limitan la conversación y ofrecen respuestas que rara vez contemplan situaciones imprevistas. El resultado es un servicio poco flexible, que no disminuye de forma sustancial la necesidad de intervención humana y que, en muchos casos, termina generando frustración en los usuarios clientes.

A su vez, se desaprovecha el potencial de los avances recientes en inteligencia artificial, particularmente los modelos de lenguaje a gran escala (LLM, por sus siglas en inglés), un conjunto de modelos de Inteligencia Artificial (IA) entrenados con grandes volúmenes de texto para comprender y generar lenguaje natural, seguir instrucciones y razonar sobre el contexto de la consulta. Estas capacidades permiten ofrecer interacciones más adaptativas y cercanas al razonamiento humano, con respuestas adecuadas al registro lingüístico y a las convenciones locales.

A medida que crece el uso de aplicaciones móviles, la banca web y los canales de mensajería, el volumen de interacciones se multiplica en los sistemas de atención. Si no existen herramientas que comprendan el contexto, utilicen el historial del cliente y propongan soluciones que den respuestas a los clientes, la consecuencia será un aumento en la carga de trabajo de los equipos humanos y una pérdida de oportunidades para mejorar la relación con clientes.

1.2 Justificación de la propuesta

La propuesta de este trabajo surge de la diferencia entre las soluciones limitadas que actualmente brindan los asistentes virtuales tradicionales implementados en Argentina y las capacidades más recientes que ofrece la tecnología. Los sistemas basados en flujos predefinidos

responden a pocas preguntas y se repiten en sus interacciones con menús de pasos simples definidos, lo que genera experiencias limitadas y, en muchos casos, insatisfacción en los clientes.

Los modelos de lenguaje a gran escala, integrados en plataformas como Salesforce, ofrecen una alternativa distinta y novedosa. A diferencia de los esquemas rígidos, pueden reconocer la intención de la consulta y dar respuestas coherentes tomando en cuenta el contexto del sistema y del usuario que está requiriendo asistencia. Esto acerca la interacción a una conversación más natural y práctica para el usuario. Además, al asumir una gran variedad de tareas, los agentes o asistentes virtuales inteligentes liberan al personal de las empresas para ocuparse de problemas más complejos y valiosos.

En el contexto del sector bancario argentino, aplicar esta tecnología puede significar más que una mejora en la atención. También implica reducir costos operativos, acelerar la resolución de consultas y fortalecer el vínculo con los clientes, uno de los principales pilares de una industria que tiene como objetivo permanente mantener una relación de confianza entre las entidades bancarias y sus clientes. Por estas razones, este trabajo se propone analizar el estado actual de la atención virtual en los bancos que operan en el país y diseñar una solución de implementación que sea realista y aplicable en el entorno financiero local.

1.3 Marco Institucional

La propuesta de intervención se orienta al sector bancario argentino, particularmente a aquellas entidades financieras que ofrecen servicios de atención al cliente a través de canales digitales. El objetivo del asistente virtual inteligente es automatizar la atención al cliente en bancos de mediana y gran envergadura que cuentan con plataformas web, aplicaciones móviles y centros de contacto en línea como parte de su estrategia de servicio al público, tanto para clientes actuales como para potenciales.

El alcance contempla la incorporación de un asistente basado en procesamiento de lenguaje natural, que se desarrolla como un agente inteligente integrado a la infraestructura tecnológica ya existente en cada banco. Para su desarrollo e implementación, se considera la plataforma Salesforce como entorno de trabajo principal, dada su capacidad de integrar datos del cliente con herramientas de inteligencia artificial.

Este proyecto está especialmente dirigido a instituciones con una marcada orientación hacia la transformación digital. Se priorizan aquellas que, aun disponiendo de cierto nivel de

automatización, enfrentan limitaciones en la experiencia del cliente y en la eficacia de sus canales de atención virtual.

1.4 Objetivos del Trabajo Final

1.4.1 Objetivo General

Diseñar e implementar un asistente virtual inteligente, basado en inteligencia artificial y procesamiento de lenguaje natural, que actúe como agente de atención al cliente en entidades bancarias argentinas.

1.4.2 Objetivos Específicos

- Analizar el estado actual de implementación de asistentes virtuales en entidades bancarias de Argentina.
- Evaluar las capacidades técnicas de la plataforma Salesforce para el desarrollo de agentes virtuales inteligentes.
- Diseñar la arquitectura de un asistente virtual basado en modelos de lenguaje a gran escala (LLM), con enfoque en atención al cliente.
- Diseñar un prototipo funcional del asistente virtual que permita evidenciar las capacidades de configuración, personalización y desarrollo en la plataforma Salesforce.

Capítulo 2 – Estado del Arte

2.1 Introducción

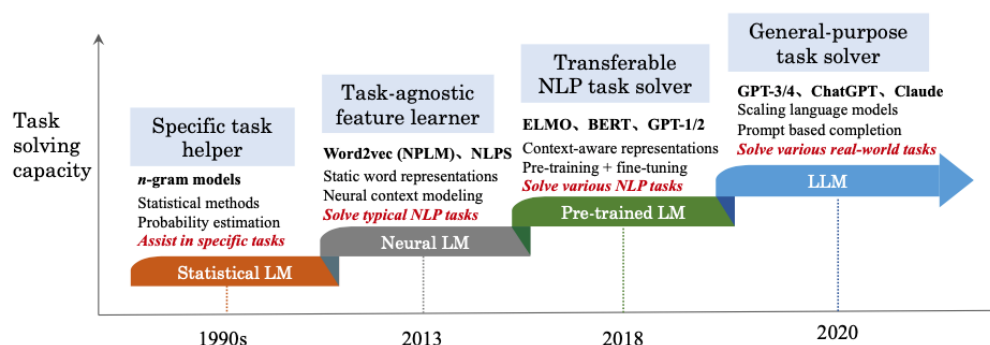
La inteligencia artificial aplicada a la atención al cliente ha sido objeto de un notable desarrollo en los últimos años, especialmente en industrias con alta demanda de interacción, como el sector financiero. En este apartado se presentan estudios recientes que demuestran cómo los modelos de lenguaje a gran escala o Large Language Models (LLM) y sus aplicaciones específicas, como los asistentes virtuales, están transformando la relación entre clientes y entidades bancarias.

Los lenguajes a gran escala constituyen una clase particular de arquitecturas de aprendizaje profundo entrenadas con grandes volúmenes de texto, capaces razonar sobre el lenguaje natural. Según Zhao et al. (2023), los LLM se refieren a modelos de lenguaje basados en la arquitectura *Transformer*, transformador, la cual puede contener cientos de miles de millones de parámetros que son entrenados con grandes volúmenes de datos textuales. De esta manera es cómo funcionan modelos como GPT-3, PaLM, Galactica y LLaMA. La gran capacidad de los LLM se basa en la comprensión del lenguaje natural de los humanos y la resolución de tareas complejas mediante generación de texto (p. 4).

La Figura 1 presenta un recorrido histórico sobre la evolución de los modelos de enfoques estadísticos centrados en tareas específicas a sistemas generales capaces de resolver problemas del mundo real en múltiples dominios.

Figura 1

Las cuatro generaciones del modelo de lenguaje desde la perspectiva de la capacidad de resolución de tareas



Nota. Obtenido de *An evolution process of the four generations of language models (LM) from the perspective of task solving capacity*, Zhao et al., 2023, A survey of large language models.

En la Figura 1 se distinguen cuatro generaciones de modelos de lenguaje. La primera corresponde a los modelos estadísticos basados en n-gramas, limitados a tareas muy acotadas mediante estimaciones probabilísticas. La segunda generación, con los modelos neuronales (Word2Vec, NPLM, entre otros), introdujo representaciones distribuidas de palabras que permitieron resolver tareas típicas de Procesamiento de Lenguaje Natural (PLN). La tercera etapa, con modelos preentrenados como ELMo, BERT o GPT-2, aportó representaciones sensibles al contexto y técnicas de *fine-tuning* (ajuste fino), lo que amplió su aplicabilidad a diversas tareas. Finalmente, a partir de 2020 surgieron los modelos de lenguaje a gran escala como GPT-3, ChatGPT o Claude, que se caracterizan por su naturaleza para solucionar tareas de propósito general, capaces de aplicar lo aprendido en un amplio rango de contextos, incluyendo servicios financieros y aplicaciones en atención al cliente. (Zhao, 2023, p. 2).

La base de los lenguajes a gran escala surge de la arquitectura Transformer, explicada por Vaswani et al. (2017) como una red formada por mecanismos de atención, prescindiendo por completo de la recurrencia (p. 1). El diseño propuesto por el autor determinó un cambio de paradigma respecto a las redes neuronales que se conocían hasta el momento, al permitir un nivel de paralelización significativamente mayor y alcanzar mejores resultados con menos tiempos de entrenamiento, ya que los Transformers permiten una paralelización significativamente mayor (Vaswani et al., 2017, p. 2). Estas ventajas son las que hicieron que los transformadores se consoliden como la arquitectura predominante en el procesamiento de lenguaje natural y que posteriormente permitiría el desarrollo de modelos de lenguajes como Generative Pre-trained Transformer (GPT), Bidirectional Encoder Representations from Transformers (BERT) o Pathways Language Mode (PaLM).

En cuanto a antecedentes recientes de asistentes virtuales, el primer caso analizado es el modelo BERTaú, desarrollado por Finardi et al. (2021), que adapta la arquitectura BERT al contexto financiero del Banco Itaú en Brasil. El modelo fue entrenado con textos reales de interacción entre clientes y el chatbot, logrando mejoras significativas: incremento del 22% en la recuperación de preguntas frecuentes, aumento del 2,1% en la detección del estado del cliente mediante análisis de sentimiento y una mejora del 4,4% en el reconocimiento de entidades nombradas. Además, se redujo en un 66% el uso de *tokens*, unidades mínimas de significado, frente a otros LLM, lo que representa un ahorro considerable de costos (p. 1). Este caso evidencia que la especialización de un modelo de lenguaje en un dominio concreto, y puntualmente en el dominio financiero, mejora la precisión y eficiencia del servicio automatizado.

En segundo lugar, Guan et al. (2023) desarrollaron un asistente virtual con capacidad para descomponer y ejecutar múltiples operaciones dentro de aplicaciones móviles, como Alipay, líder de pagos digitales en China. El sistema, basado en automatización de procesos mediante LLM, puede interpretar instrucciones complejas, generar descripciones, identificar elementos de una interfaz y anticipar errores o pasos siguientes en un proceso. En pruebas de campo, alcanzó una tasa de éxito por paso del 93,7% y por tarea completa del 70,4% (p. 5), demostrando que es posible realizar transacciones y gestiones completas sin intervención humana.

El tercer estudio, realizado por Yang et al. (2020), presenta FinBERT, un modelo entrenado específicamente para análisis de sentimiento en el ámbito financiero. A partir de un corpus de 4,9 billones de tokens compuesto por informes corporativos, transcripciones de teleconferencias y reportes de analistas, el modelo obtuvo mejoras sustanciales en métricas de clasificación de sentimiento frente a versiones genéricas de BERT, incluso con datos limitados (p. 1).

En conjunto, estos antecedentes presentan que la adaptación de modelos de lenguaje a un dominio específico como el bancario, incrementa de manera significativa la calidad de las interacciones automatizadas y la eficiencia de los procesos. Asimismo, la incorporación de automatización basada en LLM amplía las capacidades de los asistentes virtuales, posibilitando la gestión de operaciones completas y complejas sin la intervención directa de personal humano. La experiencia acumulada en mercados como Brasil y China constituye un precedente valioso para el sector bancario argentino, que aún dispone de un amplio margen para integrar este tipo de soluciones y aprovechar sus beneficios operativos y estratégicos.

2.2 Situación Actual en Bancos de Argentina

En Argentina, la banca ha experimentado una notable transformación digital en la última década. Según Galileo Financial Technologies (2025b), más del 60% de los ciudadanos utiliza servicios bancarios digitales, y las transacciones electrónicas representan más del 75% del total de operaciones. Este cambio en el comportamiento de los usuarios impulsó la adopción de herramientas de atención digital, como los chatbots, por parte de los principales bancos del país. Sin embargo, estas soluciones conversacionales siguen respondiendo a arquitecturas tradicionales basadas en reglas fijas determinadas por un árbol de decisión o procesos determinísticos, lo cual limita su capacidad para resolver situaciones complejas, adaptarse al lenguaje natural del usuario o contextualizar respuestas con conversaciones previas (p. 1).

Las limitaciones se repiten en todo el ecosistema bancario latinoamericano. Galileo Financial Technologies (2025a) también destaca que los chatbots tradicionales “no pueden comprender el contexto detrás de las interacciones con los clientes, lo que resulta en problemas no resueltos, pérdida de tiempo y una mala experiencia del usuario” (p. 1). Además, enfrentan el desafío de satisfacer a usuarios con altas expectativas de precisión, velocidad y personalización. En un entorno de creciente digitalización y necesidad de incluir más personas dentro del sistema bancario formal, se requiere un nuevo enfoque tecnológico más flexible.

Con el objetivo comprender el grado de madurez tecnológica en la atención al cliente del sector bancario argentino, se realizó un relevamiento de los asistentes virtuales implementados actualmente por distintas entidades bancarias con atención a minoristas. El análisis permitió identificar sus características principales y determinar si operan como chatbots tradicionales o como agentes virtuales con capacidades de inteligencia artificial generativa con procesamiento de lenguaje natural.

Los resultados muestran que, en la mayoría de los casos, se trata de chatbots basados en flujos conversacionales predefinidos y árboles de decisión, sin integración evidente de modelos de lenguaje a gran escala. Esto se traduce en interacciones limitadas a un conjunto de consultas frecuentes y respuestas no dinámicas y sin capacidad de adaptarse al contexto o aprender de la experiencia previa del usuario.

La Tabla 1 resume la información obtenida, detallando el nombre de los asistentes virtuales actuales y los canales de atención en los que se encuentran disponibles. Para obtener los resultados expuestos se presentan las conversaciones ejecutadas con los asistentes virtuales actuales en el Anexo 1.

Tabla 1

Chatbots de atención al cliente en bancos argentinos y sus canales disponibles

Entidad financiera	Nombre comercial	Disponibilidad de canales	Tipo de automatización
Banco Macro	BancoChat	WhatsApp	Reglas predefinidas
Banco Galicia	Gala	WhatsApp, Web	Reglas predefinidas
Banco Santander	Santi (sólo para clientes)	Web, App móvil, WhatsApp	Reglas predefinidas
Banco Comafi	Sofía	Web	PLN clásico con reglas
Banco Hipotecario	El Búho	Web, WhatsApp	Reglas predefinidas
Banco de la Ciudad de Buenos Aires	Bit	WhatsApp	Reglas predefinidas
Banco de la Provincia de Buenos Aires	Asistente Virtual	Web	PLN clásico con reglas
Banco Patagonia	Padi	Web, App móvil	Reglas predefinidas
Banco Supervielle	Supervielle Chat	WhatsApp	Reglas predefinidas

Nota. Elaboración propia. El análisis se realizó a partir de la exploración directa de los canales oficiales de atención de cada banco elegido.

El relevamiento presenta que los bancos argentinos han avanzado en la adopción de herramientas digitales para la atención al cliente, pero a su vez el modelo predominante continúa siendo el de chatbots con respuestas predefinidas o con ciertas limitaciones en cuanto a la automatización de procesos expuestos al cliente. El escenario actual representa una oportunidad para evolucionar hacia agentes virtuales inteligentes, capaces de interpretar la intención de las consultas y ofrecer interacciones más dinámicas, precisas y personalizadas, alineadas con las tendencias internacionales descritas anteriormente.

En Argentina, uno de los antecedentes más significativos en automatización conversacional fue alcanzado por el Banco Comafi, que en 2019 implementó su asistente virtual denominado Sofía, desarrollado en colaboración con la empresa tecnológica Aivo.

Desde entonces este chatbot ha sido mejorado para lograr respuestas lo más cercano posible a un lenguaje humano con procesamiento de lenguaje natural clásico, aunque por el momento con ciertas limitaciones en los tipos de respuestas, ya que no se utiliza estrictamente un LLM como soporte para un entendimiento más profundo del lenguaje y la búsqueda de respuestas en archivos. Según Aivo (2025), Sofia procesa más de 100.000 consultas mensuales, con un promedio de respuesta de dos minutos y medio y un nivel de efectividad del 98% (párr. 1).

Banco Santander anunció en agosto de 2025 que su estrategia “data & AI-first” se consolidó a partir de un acuerdo de colaboración con OpenAI, empresa responsable del desarrollo de GPT, con el objetivo de avanzar hacia un modelo de banco “nativo en IA”. Según Ricardo Martín Manjón, responsable global de Datos e IA de Santander, “la ventaja no está en ‘usar la inteligencia artificial’ sin más, sino en integrarla en la forma de trabajar” (Santander, 2025). La entidad bancaria ha desplegado pilotos internos de inteligencia artificial en diversas áreas y reporta que más del 40% de las interacciones en sus centros de contacto ya cuentan con apoyo de estas herramientas, lo que se traduce en importantes ahorros y eficiencia operativa. Según Manjón, la alianza con OpenAI no solo busca optimizar procesos, sino también “impulsar un cambio en cultura, productividad y experiencia”, integrando agentes inteligentes en múltiples funciones y construyendo un modelo operativo basado en datos confiables y tecnología flexible (Santander, 2025). Estas implementaciones están impactando principalmente a Santander España con proyección de expandir los avances a la Argentina a mediados de 2026.

También existen bancos argentinos que están explorando casos puntuales de inteligencia artificial generativa. Un caso reciente es el del banco BBVA, que incorporó herramientas como Gemini, el motor de IA de Google, y ChatGPT Enterprise (la versión corporativa del modelo de OpenAI) para fortalecer procesos internos y desarrollar asistentes que asisten a sus equipos en tareas de soporte y programación. El banco Supervielle, por su parte, inició este mismo camino con un asistente disponible en WhatsApp basado en menús predefinidos y actualmente experimenta con proyectos de IA contextual capaces de interpretar el historial y las necesidades del usuario para ofrecer respuestas más personalizadas (Ámbito, 2025).

La banca argentina cuenta con capacidades técnicas y un ecosistema de proveedores que permiten anticipar un desarrollo futuro de agentes con inteligencia artificial generativa, pero aún no se ha producido una adopción consolidada en un caso empírico que se encuentre en funcionamiento desplegado a canales digitales para clientes. Esto confirma la oportunidad

de innovar con propuestas como la presente que propone diseñar un terreno de implementación viable para el sector financiero nacional.

2.3 Implementación de LLM en Servicios Financieros

2.3.1 Casos Internacionales del Dominio Financiero

Uno de los casos más destacados a nivel latinoamericano es el del Banco de Chile, que en 2023 presentó a FANi, su asistente virtual integrado en su ecosistema digital. Este asistente fue diseñado para atender consultas frecuentes y acompañar a los usuarios durante los procesos de apertura y administración de cuentas digitales, contribuyendo a una experiencia de autoservicio dentro del portal oficial del banco, y que este año incorporó también algunas interacciones con procesamiento de lenguaje natural (Banco de Chile, s. f.). La implementación de FANi fue desarrollada por la consultora Apiux, utilizando infraestructura sobre Google Cloud y el servicio Dialogflow CX. Según Apiux (2025), el asistente fue capaz de sostener conversaciones contextuales, integrarse con sistemas bancarios centrales y mejorar la eficiencia del proceso de apertura de cuentas.

A nivel global, la implementación de modelos de lenguaje a gran escala en el sector financiero también ha mostrado resultados significativos cuando estos se entrenan con datos especializados, marcando un precedente importante para motivar a los bancos a incorporar inteligencia artificial para mejorar la ejecución de procesos. Un caso destacado es BloombergGPT, un modelo de 50 billones de parámetros entrenado con un corpus mixto de datos financieros y generales, que superó ampliamente a modelos generales en tareas como respuestas a preguntas, reconocimiento de entidades y análisis de sentimiento orientados a finanzas. Este enfoque presenta que, en contextos de alta complejidad como el financiero, adaptar un LLM al lenguaje y a las operaciones del sector potencia su utilidad para tareas críticas (Wu et al., 2023, pp. 4, 25).

Otra aproximación reciente es la de FinGPT, un proyecto de código abierto que facilita a instituciones financieras el desarrollo de LLM propios mediante limpieza de datos y técnicas de ajuste fino eficientes, como LoRA. Este marco permite construir asistentes financieros (por ejemplo, para asesoramiento, análisis documental o automatización de procesos) con control total sobre la procedencia de los datos y las reglas de entrenamiento, aspecto clave para cumplir con normativas de privacidad y auditoría (Yang et al., 2023, p. 2). En este caso se buscó un enfoque más centrado en los datos, lo que ofrece a las entidades la posibilidad de experimentar

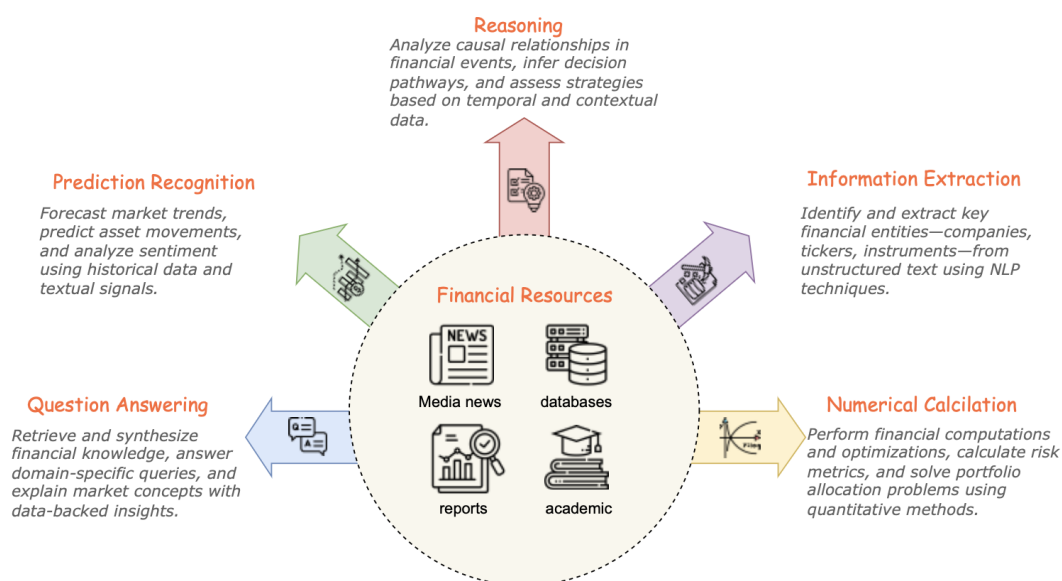
con aplicaciones de inteligencia artificial generativa sin depender de proveedores cerrados, manteniendo transparencia y capacidad de auditoría en cada etapa (Yang et al., 2023, p. 4).

Tras las iniciativas de código abierto, en 2025 se propuso también una investigación conocida como BizFinBench, considerado el primer *benchmark* (prueba de referencia) específicamente orientado a evaluar modelos de lenguaje en aplicaciones financieras reales. Como señalan sus autores, BizFinBench fue el primer conjunto de evaluación diseñado específicamente para medir modelos de lenguaje en aplicaciones financieras con casos reales (Lu et al., 2025, p. 1). El *dataset* (conjunto de datos) incluyó 6.781 consultas, distribuidas en nueve categorías y agrupadas en cinco dimensiones principales: cálculo numérico, razonamiento, extracción de información, reconocimiento de predicciones y preguntas basadas en conocimiento (Lu et al., 2025, p. 3).

En la siguiente figura se muestra cómo se distribuyen las tareas de BizFinBench a lo largo de las cinco dimensiones clave. Cada una de ellas está orientada a evaluar una capacidad específica de los modelos de lenguaje aplicados al dominio financiero. Además, se sintetiza el enfoque central que caracteriza a cada dimensión, lo que permite visualizar de manera esquemática el alcance del benchmark (Lu et al., 2025, p. 4).

Figura 2

Distribución de tareas en BizFinBench según cinco dimensiones clave



Nota. Obtenido de *Distribution of tasks in BizFinBench across five key dimensions* por Lu et al., 2025, BizFinBench: A Business-Driven Real-World Financial Benchmark for Evaluating LLMs.

Estas cinco dimensiones observadas en la Figura 2 reflejan las competencias que los modelos de lenguaje deben incluir en el ámbito financiero: desde responder preguntas específicas y extraer información relevante, hasta realizar cálculos numéricos y desplegar razonamiento aplicado a contextos de negocio. La clasificación propuesta por BizFinBench permite evaluar no solo la exactitud de las respuestas, sino también la capacidad de los modelos para enfrentar la complejidad de escenarios reales en servicios financieros, aportando un marco de referencia útil para el análisis de futuras implementaciones (Lu et al., 2025, p. 4).

Uno de los hallazgos más relevantes es que, si bien los modelos más avanzados alcanzan un buen desempeño en tareas específicas, no existe aún un modelo que pueda garantizar un dominio completo en todas las tareas (Lu et al., 2025, p. 1), lo que evidencia que la fiabilidad de los LLM financieros aún depende de la naturaleza de la tarea. Además, los autores destacan que este benchmark fue construido sobre operaciones y escenarios de negocio, diferenciándose de aproximaciones previas al estar diseñado explícitamente en torno a operaciones financieras reales e interacciones con usuarios (Lu et al., 2025, p. 4).

Bigeard et al. (2025) también presenta Finance Agent Benchmark (FAB), un marco de evaluación orientado a medir el desempeño de agentes financieros basados en LLM frente a problemas de investigación reales. Tal como explican sus autores, el punto de referencia estudiado incluye problemas de investigación financiera, los cuales requieren que los modelos de lenguaje realicen un análisis complejo utilizando presentaciones recientes ante la Comisión de Bolsa y Valores de Estados Unidos (SEC, por sus siglas en inglés). A diferencia de otros benchmarks centrados en tareas aisladas, este enfoque evalúa flujos completos de trabajo (end-to-end), desde la búsqueda y recuperación de información hasta el razonamiento y la elaboración de respuestas.

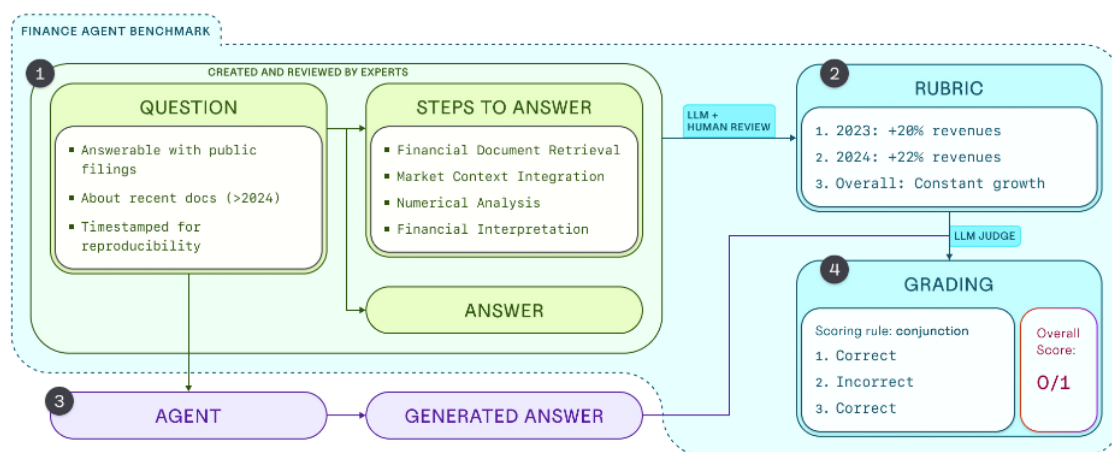
El proceso de evaluación del FAB se organiza en cuatro etapas encadenadas: (1) *Data Creation* (creación de datos), etapa en la que los expertos del sector formulan preguntas prácticas y fijan respuestas de referencia a partir de documentos públicos recientes; (2) *Rubric Development* (elaboración de rúbricas), que convierte ese material experto en rúbricas con pasos de cálculo y de razonamiento esperados para la evaluación sistemática; (3) *Agent Evaluation* (evaluación de agentes), en la que las preguntas se ejecutan mediante agentes LLM provistos de herramientas para producir respuestas; y (4) *Answer Grading* (calificación de respuestas), que aplica un esquema de LLM-as-a-Judge con reglas de conjunción para decidir la corrección en múltiples criterios. El método LLM-as-a-Judge se basa en utilizar un modelo de lenguaje de gran escala como evaluador automático de las respuestas generadas por otros

modelos. En la práctica, el LLM actúa como un juez que compara la salida de un agente con la respuesta de referencia y determina si cumple con los criterios fijados en la rúbrica. El enfoque permite calificar de manera sistemática grandes volúmenes de respuestas sin necesidad de revisión humana en cada caso. Sin embargo, los autores advierten que este mecanismo puede introducir sesgos propios del modelo evaluador, motivo por el cual en el FAB se emplean reglas de conjunción y verificaciones adicionales para aumentar la fiabilidad del proceso (Bigéard et al., 2025, p. 2).

Este flujo se resume en la siguiente figura, que presenta de manera esquemática la metodología aplicada en el benchmark de los autores:

Figura 3

Arquitectura del análisis del Agente financiero



Nota. Obtenido de *Architecture of the Finance Agent Benchmark* (p. 2) por Bigéard et al., 2025, Finance Agent Benchmark: Benchmarking LLMs on Real-world Financial Research Tasks.

En el diagrama anterior se presenta cómo el proceso comienza con la ejecución de una pregunta y la definición de pasos para responderla, seguido del desarrollo de rúbricas de evaluación, la ejecución de agentes equipados con herramientas externas a su propio funcionamiento y, finalmente, la calificación automática de respuestas mediante el enfoque LLM-as-a-Judge. La Figura 3 demuestra que el benchmark no se limita a evaluar tareas puntuales que los agentes deben resolver, sino que reproduce un flujo de investigación financiera de extremo a extremo, acercándose a la lógica operativa que se requeriría de una persona que actúe como asistente bancario.

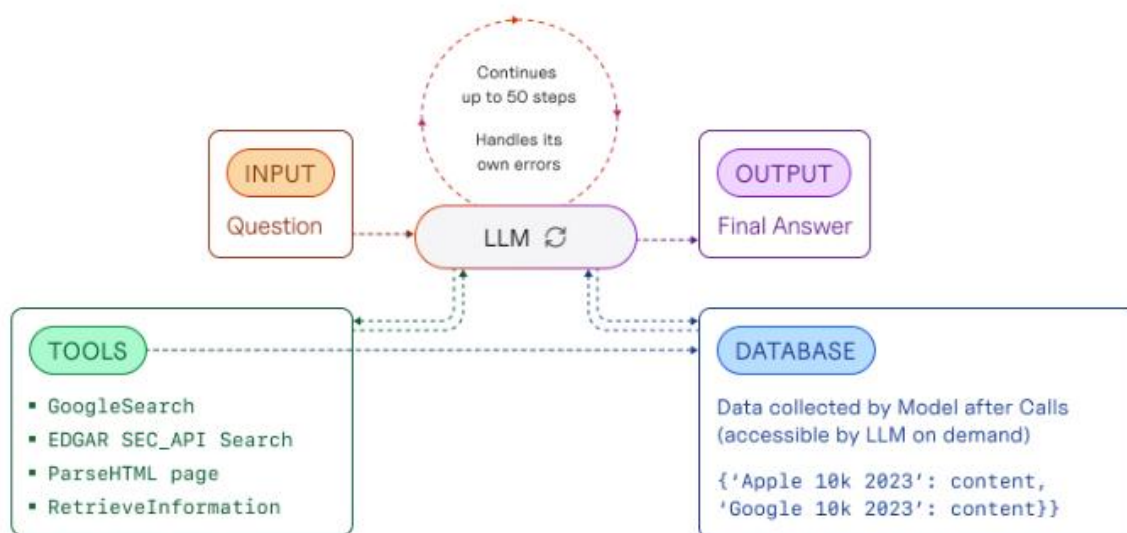
El dataset construido para el FAB se formó de un total de 537 preguntas distribuidas en nueve categorías de tareas, que van desde la recuperación simple de información hasta actividades de modelado financiero y análisis de mercado. Cada entrada incluye no solo la pregunta y su respuesta esperada, sino también los documentos fuente necesarios y un enfoque paso a paso de la solución, lo que permite evaluar con transparencia los procesos de razonamiento seguidos por los agentes. Con el fin de asegurar la calidad, cada pregunta fue sometida a una revisión cruzada por un experto distinto al autor original, verificando los cálculos y la validez del contenido antes de su incorporación definitiva al benchmark (Bigeard et al., 2025, p. 3).

Como lo indica Bigeard et al. (2025, p. 5), el FAB incorpora además el entorno agentico “agent harness”, que simula el trabajo de un empleado real al equipar a los modelos con herramientas externas para resolver las consultas. Este entorno incluye funciones de herramientas como Google Search (buscador web) y EdgarSearch (repositorio de presentaciones regulatorias de la Comisión de Bolsa y Valores de Estados Unidos), la extracción de texto en formato HTML y un motor de recuperación documental especializado. De esta forma, los agentes no se limitan a responder con la información almacenada en sus parámetros, sino que pueden consultar otras fuentes para después razonar con evidencia actualizada al momento de resolver los problemas. Este diseño aproxima la evaluación a un escenario corporativo real, donde los asistentes deben interactuar con múltiples fuentes heterogéneas y validar sus respuestas antes de entregarlas al usuario (Bigeard et al., 2025, p. 2).

La siguiente imagen ilustra la arquitectura del entorno propuesto por Bigeard et al. (2025) y las herramientas que se integran en el flujo de trabajo de los agentes.

Figura 4

Arquitectura del agente financiero



Nota. Obtenido de *Architecture of the Finance Agent Benchmark* (p. 2) por Bigeard et al., 2025, Finance Agent Benchmark: Benchmarking LLMs on Real-world Financial Research Tasks

La Figura 4 permite observar cómo el agente financiero no actúa de manera aislada, sino que se encuentra inmerso en un ecosistema de herramientas que amplían sus capacidades. Cada módulo dentro del recuadro herramientas representa una función crítica: desde la búsqueda en línea y el acceso a repositorios regulatorios, hasta la extracción de datos estructurados que luego son interpretados por el modelo. De esta forma, el benchmark no evalúa únicamente la calidad del texto generado, sino la capacidad del agente para orquestar distintas fuentes de información y utilizarlas en un flujo de trabajo coherente, lo que se acerca mucho más a las exigencias de un asistente bancario real. En la base de datos se encuentran los datos estructurados de las empresas utilizadas y el modelo LLM funciona repetidamente generando una respuesta final para responder a una pregunta que nació como punto de ingreso al procesamiento, en la Figura 4 indicado como “input”. IBM (s.f.a) define a los datos estructurados como datos que “se organizan en un formato claro y predefinido” y pueden incluir datos cuantitativos como el valor de un precio de un producto o datos cualitativos representados en ocasiones como fechas, nombres o direcciones. Es decir que un informe con columnas ordenadas y filas se consideran datos estructurados. En cambio, los datos no estructurados según IBM (s.f.a) son aquellos que “no tienen un formato predefinido”, lo que quiere decir que son representados por correos electrónicos, archivos en formato PDF, documentos de texto,

archivos de mensajes, imágenes, videos o incluso datos de sensores de dispositivos. La gran habilidad que introducen los LLM en la IA es que no sólo pueden interpretar datos estructurados, sino que también pueden comprender y procesar datos no estructurados.

En términos de desempeño, los autores destacan que el modelo con mejor rendimiento fue el o3 con un 46,8% de precisión, y con un costo promedio de 3,79 dólares por consulta y un tiempo de 3,1 minutos por respuesta (Bigeard et al., 2025, p. 9). Al contrastar estos resultados con la labor de expertos humanos, las personas expertas en finanzas que pueden realizar las mismas tareas requieren 16,8 minutos de resolución por respuesta y un costo de 25,66 dólares por consulta (Bigeard et al., 2025, p. 3). Esto significa que, aunque los modelos no alcanzan un nivel de precisión suficiente para un despliegue autónomo en tareas críticas, sí ofrecen ventajas claras en términos de tiempo de respuesta y costos ante esquemas de trabajo tradicionales, funcionando mejor como asistentes que aumentan la productividad que como sustitutos directos.

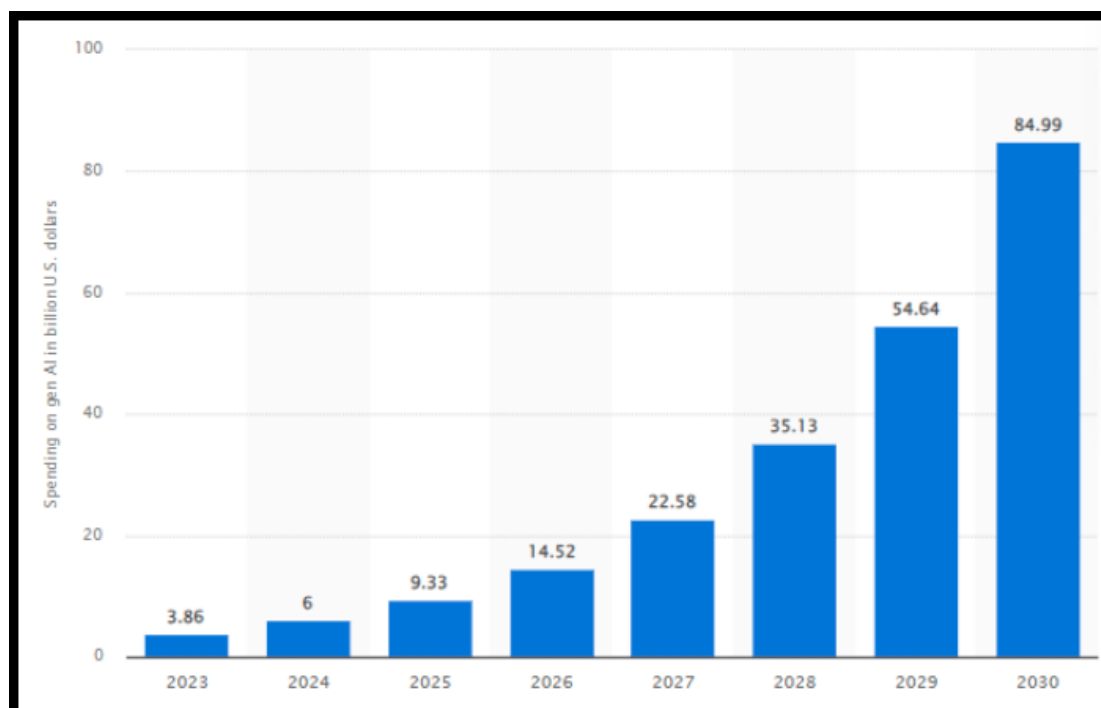
2.3.2 Inteligencia Artificial en Atención al Cliente

La inteligencia artificial se ha consolidado como un recurso estratégico para mejorar la atención al cliente en el sector financiero con la creación de asistentes de atención que han ido adquiriendo más autonomía. A su vez, la aplicación de otras capacidades de IA en los bancos ha ido aumentando en los últimos años y se espera que el proceso de automatización de tareas se profundice más aún. Ante este escenario de una demanda creciente del sector bancario en la utilización de la tecnología automatizada, Todupunuri (2024) analiza que esto ha impulsado el uso de asistentes virtuales y chatbots potenciados por inteligencia artificial para ofrecer a los clientes una solución de soporte 24/7. El objetivo es brindar soluciones a consultas rutinarias, reduciendo la carga de trabajo de las personas que se desempeñan asociados a la industria bancaria (p. 206).

En la Figura 5, se presenta el gasto en inversiones de inteligencia artificial en el sector bancario.

Figura 5

Gasto en inteligencia artificial (IA) en la industria bancaria a nivel mundial



Nota. Obtenido de *Artificial intelligence (AI) spending on the banking industry globally* (p. 206) por Todupunuri, 2024, *The Future of Conversational AI in Banking: A Case Study on Virtual Assistants and Chatbots: Exploring the Impact of AI-Powered Virtual Assistants on Customer Service Efficiency and Satisfaction*

Todupunuri (2024) afirma que los chatbots se han convertido en partes integrales del sector bancario porque contribuyen a la eficiencia operativa y a la reducción de costos (p. 210). En el mismo trabajo se destaca que los asistentes virtuales son soluciones altamente escalables durante las horas pico o ante incrementos inesperados en las consultas de los clientes (p. 207), lo que evidencia su potencial para absorber grandes volúmenes de interacciones en momentos críticos. El autor indica que los asistentes virtuales utilizan analítica de datos para ofrecer servicios personalizados a los consumidores en el sector bancario (p. 211), una funcionalidad que amplía el valor estratégico de estas tecnologías más allá de la simple automatización. Todupunuri (2024) también advierte que la IA conversacional incrementó las preocupaciones sobre la privacidad y seguridad de los datos en el sector bancario (p. 211), lo cual exige que cualquier implementación debe acompañarse de marcos de gobernanza de datos y medidas de ciberseguridad robustas.

El uso de IA en banca no se limita al contacto directo con los clientes, sino que también mejora el trabajo interno de los empleados. Morgan Stanley, por ejemplo, viene desarrollando herramientas internas basadas en modelos como GPT-4 para asistir a sus equipos en la síntesis de información, respuestas a consultas frecuentes y transcripción de llamadas. Según Business Insider (2025), la empresa estadounidense lanzó su primera versión de inteligencia artificial en septiembre de 2023 para que sea utilizada por asesores financieros y personal de soporte que ayudan a los clientes a gestionar su dinero ofreciendo respuestas fundamentadas en la biblioteca de la empresa, que contiene más de 100.000 informes y documentos de investigación (párr. 13).

JPMorgan Chase, en la misma línea, ha desarrollado un conjunto interno de modelos de lenguaje (LLM Suite) utilizado por más de 200.000 empleados. En una entrevista publicada por The Wall Street Journal (2025), Teresa Heitsenrether, directora de Análisis de Datos de JPMorgan Chase, propone pensar cuando un cliente que llama al banco, podría hacerlo para consultar sobre su préstamo para automóvil, de su hipoteca o sobre su cuenta de depósito o su tarjeta de crédito, por lo que la persona al otro lado del teléfono debe ser capaz de manejar todos esos productos diferentes. Heitsenrether (2025, citada en The Wall Street Journal, 2025), sostiene que los modelos permiten que el agente del centro de atención al cliente encuentre la información más rápido y responda a las preguntas de manera más rápida y precisa (p. 1).

De acuerdo con la empresa OpenText (2024), la IA está poniendo a la banca en transición desde un mundo de flujos de trabajo y procesos hacia un mundo de inteligencia y automatización (p. 2). La implementación de esta tecnología permite a las entidades financieras analizar grandes volúmenes de datos en tiempo real, identificar patrones y generar perspectivas de negocio que son fundamentales para desarrollar experiencias personalizadas, detectar fraudes avanzados y mejorar la eficiencia operativa (p. 3). Estos avances confirman que la IA ya no es una herramienta auxiliar, sino un elemento estratégico clave para la modernización de la atención bancaria.

2.4 Salesforce

2.4.1 Evolución de la Plataforma hacia Agentforce

Salesforce, empresa líder mundial en soluciones de gestión de relaciones con clientes, más conocido como CRM (Customer Relationship Management, en inglés), ha consolidado su posición de dominio en el mercado gracias a su innovación tecnológica y su visión centrada en el cliente (Salesforce, 2024c).

De esta manera, en la Figura 6 se observa que la empresa fundada en San Francisco en 1999 tiene algunas ventajas sobre otras tecnologías como líder en la industria tecnológica, lo que la convierte en una gran oportunidad para desarrollar asistentes virtuales que puedan trabajar con información actualizada de clientes en tiempo real.

Figura 6

Cuadrante Mágico para Plataformas de Automatización de la Fuerza de Ventas



Nota. Obtenido de *Magic Quadrant for Sales Force Automation Platforms* (p. 1) por Salesforce, 2024c, Salesforce Named a Leader in the 2024 Gartner® Magic Quadrant™ for Sales Force Automation Platforms

Salesforce ha desempeñado un papel destacado en la transición global de los chatbots tradicionales hacia agentes virtuales autónomos, concepto que la empresa denomina Agentforce. La presentación oficial de Agentforce ocurrió en septiembre de 2024, durante la conferencia Dreamforce, donde Salesforce lo definió como la IA que permite a las empresas impulsar el éxito del cliente, brindando a las organizaciones agentes autónomos que aumentan

la eficiencia y la satisfacción en el servicio, las ventas, el marketing y el comercio (Salesforce, 2024d, párr. 1).

Un mes después, en octubre de 2024, la plataforma alcanzó su disponibilidad general para todos sus clientes, permitiendo a las organizaciones desplegar agentes confiables y autónomos a gran escala (Salesforce, 2024e, p. 1).

En diciembre de 2024, Salesforce comunicó la disponibilidad de Agentforce 2.0, con capacidades mejoradas para personalizar interacciones, ampliar integraciones con sistemas externos y optimizar flujos de trabajo (Salesforce, 2024b, p. 1). En marzo del mismo año, la versión Agentforce 2dx incorporó funcionalidades de razonamiento proactivo, soporte multimodal (procesamiento de archivos) e integración directa con herramientas como Slack y Mulesoft, lo que facilitó la creación de experiencias unificadas en diversos canales corporativos. Slack es una plataforma de comunicación pensada para la colaboración entre personas de una empresa, mientras que Mulesoft es un sistema dedicado a la construcción, mantenimiento y monitoreo de integraciones entre sistemas diversos (Salesforce, 2025i, p. 1). Finalmente, en junio de 2025, Salesforce presentó Agentforce 3.0, enfocada en resolver las principales barreras para la adopción masiva de agentes de IA: visibilidad y control. Esta última versión añadió un centro de monitoreo, más de cien acciones prediseñadas y protocolos de integración que permiten escalar agentes en entornos complejos (Salesforce, 2025j).

La evolución de Agentforce presenta una estrategia de desarrollo progresivo orientada a consolidar agentes virtuales plenamente autónomos, capaces de integrarse en múltiples sistemas empresariales y operar con altos niveles de fiabilidad. Estas características la convierten en una plataforma especialmente adecuada y actualizada para implementar soluciones avanzadas de atención al cliente en sectores como el bancario, donde la velocidad de respuesta, la precisión y la personalización son factores críticos para la experiencia del usuario.

2.4.2 Agentforce para Servicios Financieros

Salesforce (2025e) ha presentado algunos casos de uso nativos que demuestran la capacidad operativa de Agentforce en servicios bancarios. Por ejemplo, el caso “Denunciar tarjeta perdida” permite generar automáticamente un registro con los datos del cliente y crear un caso de reemplazo, agilizando la resolución de la solicitud (párr. 2). El flujo “Reversión de tarifas” automatiza el proceso de devolución de cargos no reconocidos a partir de datos de cuentas financieras y transacciones (párr. 6), mientras que el tema “Detener el pago de un cheque” posibilita completar solicitudes automáticas sin intervención manual (párr. 10).

Asimismo, el caso “Notificación de viaje” integra información de distintos sistemas para activar configuraciones de uso internacional de tarjetas, reduciendo los errores y tiempos de espera (párr. 14).

Estos procesos utilizan objetos de datos específicos del modelo bancario de Salesforce, como cuenta financiera o transacción, lo que permite mantener trazabilidad y cumplimiento regulatorio, y constituyen un antecedente técnico para la implementación de asistentes virtuales personalizados basados en procesamiento con LLM.

2.4.3 Acceso a Datos Centralizados

Para que Agentforce pueda funcionar como un asistente virtual inteligente, requiere datos que puedan ser utilizados para comprender el contexto en el cual se desarrolla. La centralización de la información del cliente en una única plataforma reduce la fragmentación de datos, habilita perfiles unificados y estandariza el acceso para todos los equipos de la organización.

En el ecosistema de Salesforce (2025h), Data Cloud armoniza datos heterogéneos hacia un modelo común y unifica identidades mediante reglas de coincidencia y reconciliación, lo que permite construir una Vista 360° consistente para que los agentes de IA puedan funcionar con información fidedigna de los clientes. Esta consolidación del perfil del cliente no es solamente operativa, sino que también facilita la toma de decisiones y la personalización a escala, al evitar duplicados y sesgos de origen de datos, y al mismo tiempo mantiene controles de gobierno y seguridad sobre datos estructurados y no estructurados, requisito crítico en instituciones financieras. Data Cloud no se limita a integrar datos estructurados, sino que también incorpora información no estructurada y metadatos específicos del negocio. La plataforma ofrece reglas de coincidencia exacta o difusa y procesos de reconciliación, lo que posibilita administrar múltiples conjuntos de reglas para resolver identidades de manera flexible. De este modo, se logra una visión integral que alimenta tanto la analítica avanzada como la personalización en tiempo real. Según Salesforce (s. f.d), el valor de esta arquitectura radica en que cada interacción puede apoyarse en métricas derivadas, como el valor del ciclo de vida del cliente, calculadas a partir de los datos unificados. Estas capacidades permiten segmentar clientes, anticipar comportamientos y activar experiencias de múltiples canales con información gobernada de extremo a extremo. Data Cloud también aplica un enfoque robusto para resolver identidades mediante conjuntos de reglas (Identity Resolution Rulesets) que se combinan con reglas de emparejamiento (match rules) y reglas de reconciliación. Las primeras pueden configurarse como exactas, es decir que requieren coincidencia literal de valores (un

correo electrónico o un número de teléfono), o como difusas, que permiten asociar entradas similares aun con diferencias de escritura, como nombres abreviados o con errores tipográficos. Esta flexibilidad posibilita una unificación precisa o más inclusiva según la necesidad del negocio (Salesforce, s. f.c). Una vez que varios registros se consideran equivalentes, las reglas de reconciliación definen qué valor se conserva en el perfil consolidado, pudiendo priorizar, por ejemplo, el dato más reciente, el más frecuente o aquel que proviene de una fuente determinada (Salesforce, s. f.d).

Enfocado en servicios financieros, Salesforce también cuenta con un producto específico denominado Financial Services Cloud (FSC) que aporta un modelo de datos específico para personas, hogares y relaciones financieras, ordenando las entidades (clientes, cuentas financieras, grupos de relación, entre otras) y sus vínculos de manera nativa en la misma plataforma en la que se ejecutan los procesos liderados por IA. Esto habilita perfiles completos del cliente y de su grupo familiar para que las empresas puedan realizar una gestión coherente de interacciones, metas y actividad histórica en una única vista de sus clientes (Salesforce, 2025g, p. 9).

Adicionalmente, Salesforce (2025g, p. 37) detalla que el producto de industria FSC incorpora además otro producto llamado CRM Analytics con tableros prediseñados para banca minorista, patrimonial y seguros, que integran indicadores clave. Estos tableros de mando potencian la capacidad de los asesores para identificar oportunidades, medir resultados y tomar decisiones basadas en datos, reforzando la lógica de un acceso centralizado a la información.

Capítulo 3 – Marco Teórico

3.1 Inteligencia Artificial

3.1.1 Introducción

El desarrollo de soluciones basadas en inteligencia artificial ha transformado los modelos de atención al cliente en distintos sectores económicos. En particular, el uso de asistentes virtuales ha evolucionado desde simples interfaces de preguntas y respuestas hacia agentes inteligentes capaces de interpretar lenguaje natural y adaptarse a diversos contextos de conversación. Esta evolución ha sido posible gracias al avance de los LLM que permiten simular interacciones humanas con altos niveles de comprensión y precisión.

La inteligencia artificial generativa, conocida como GenAI, se basa en modelos de aprendizaje automático y, en particular, en redes neuronales profundas como las redes generativas antagónicas (GAN) y los LLM. Los LLM (como GPT, Claude o PaLM) representan un salto cualitativo en la capacidad de la IA para procesar y generar lenguaje humano. Estos modelos son capaces de comprender contextos complejos, mantener coherencia en largas conversaciones y responder a instrucciones con un nivel alto de pertinencia. De acuerdo con Bommasani et al. (2021, p. 4), los LLM están contruidos sobre arquitecturas de transformadores, que permiten paralelizar el entrenamiento y optimizar la comprensión del significado contextual en grandes volúmenes de texto.

Una de las principales ventajas de la inteligencia artificial generativa en entornos profesionales es su capacidad de adaptación. En lugar de depender de reglas fijas, los modelos generativos aprenden patrones desde el análisis de los datos, lo que les permite responder a una amplia variedad de situaciones sin necesidad de una programación específica para cada caso. Según Chen et al. (2023, p. 10), la GenAI puede proporcionar soluciones automatizadas de atención al cliente y ayudar a las empresas a aumentar la eficiencia, reducir costes y mejorar la experiencia del cliente, a partir del entendimiento de las necesidades del cliente y la personalización de sus estrategias de marketing.

En el ámbito empresarial, plataformas como Salesforce han comenzado a integrar modelos generativos para potenciar sus asistentes virtuales y mejorar la experiencia de usuario. Esto representa una evolución respecto de los antiguos chatbots que operaban con reglas estáticas. De este modo se lanzó el lanzamiento de Agentforce, combinando datos de clientes con capacidades generativas para crear interacciones más naturales y precisas. Según Salesforce (s.f.a), Agentforce es una aplicación de inteligencia artificial proactiva y autónoma

que se utiliza para brindar soporte especializado a empleados de compañías y clientes en general. Valoir (2025, p. 2) asegura que Agentforce brinda a las organizaciones la posibilidad de desplegar agentes autónomos de inteligencia artificial garantizando un promedio que es 16 veces más rápido que otros enfoques y aumentando en un 75% la precisión de las respuestas.

3.1.2 Fundamentos de la Inteligencia Artificial

Los orígenes de la inteligencia artificial moderna se remontan al trabajo pionero de Alan Turing, matemático inglés que en 1950 planteó la pregunta: ¿Pueden pensar las máquinas? (Turing, 1950, p. 433), en su trabajo sobre computación que planteaba el denominado “The imitation game”, hoy conocido como Test de Turing, en el que un interrogador debe distinguir entre respuestas de una persona y las de una máquina. Turing afirmaba que una forma de pensar este problema es hacerlo en términos de un juego en el que se necesitan tres personas: un hombre (A), una mujer (B) y un interrogador (C) que puede ser de cualquier género. El interrogador permanece en una habitación separada de los otros dos y el objetivo del juego es que el interrogador logre determinar cuál de los otros dos es el hombre y cuál es la mujer a partir de lo que va conociendo de ambos. Este planteamiento explicado en la obra de Turing constituyó un hito fundacional para la computación como campo de estudio, ya que sentó las bases de cómo evaluar el comportamiento inteligente de los sistemas, marco que hoy se refleja en la manera en que se construyen y prueban los asistentes virtuales.

En 1955 se logró comenzar a construir las bases científicas de lo que hoy se conoce como el campo de la inteligencia artificial. Ese año, John McCarthy, Marvin Minsky, Nathaniel Rochester y Claude Shannon redactaron la propuesta de la Conferencia de Dartmouth, considerada el acto fundacional de la disciplina. Los autores propusieron realizar un estudio de dos meses con diez investigadores sobre inteligencia artificial durante el verano de 1956 en Dartmouth College, en Hanover, New Hampshire, Estados Unidos. El estudio se realizó con el objetivo de asumir que todo aspecto del aprendizaje o cualquier otra característica de la inteligencia puede, en principio, describirse con tanta precisión que una máquina pueda simularlo (McCarthy et al., 1956, p. 2). Esta declaración sintetiza la hipótesis fundacional de la IA en la que la inteligencia humana puede formalizarse en descripciones precisas y luego implementarse en máquinas capaces de reproducirla.

Los autores (McCarthy et al., 1956, p. 2) también definieron una serie de aspectos concretos que debían abordarse para avanzar en el estudio de la inteligencia artificial. Entre ellos, destacaron la capacidad de los computadores automáticos para ejecutar tareas que hasta ese momento solo realizaban personas, y la posibilidad de programar un lenguaje que

permitiera simular el razonamiento humano. Plantearon la necesidad de investigar sobre redes de neuronas como base para la representación de conceptos, y de desarrollar teorías acerca de la dimensión y eficiencia de los cálculos, con el fin de medir y optimizar la complejidad de las operaciones. Otro punto señalado fue la importancia de la automejora, es decir, que una máquina pudiera aprender de su propia experiencia para incrementar su rendimiento. También consideraron relevante el estudio de distintos niveles de abstracción para clasificar y organizar datos sensoriales, así como la inclusión de azar y creatividad como elementos que distinguen el pensamiento humano innovador del meramente repetitivo.

La inteligencia artificial es un campo dedicado no solo a comprender, sino también a construir entidades inteligentes capaces de actuar de manera efectiva y segura en una amplia variedad de situaciones. Como indican Russell y Norvig (2021, p. 19), la IA se ocupa de construir entidades inteligentes, es decir, máquinas que pueden calcular cómo actuar de manera efectiva y segura en una amplia variedad de situaciones, simulando de alguna manera al funcionamiento humano. Por su parte, McCarthy (2004) define a la inteligencia artificial como la ciencia e ingeniería de crear máquinas inteligentes, especialmente programas informáticos inteligentes que comprendan la inteligencia humana, pero no sólo limitado a métodos que sean biológicamente observables (p. 2).

Dentro de todo el espectro de la inteligencia artificial, se encuentra el subcampo Lenguaje Artificial (Language AI), que se enfoca en desarrollar tecnologías capaces de comprender, procesar y generar lenguaje humano, siendo muchas veces utilizado como sinónimo de procesamiento de lenguaje natural. Como explican Alammari y Grootendorst (2024, p. 20), el lenguaje artificial se centra en desarrollar tecnologías capaces de entender, procesar y generar lenguaje humano.

A medida que se comenzaron a debatir los alcances de la inteligencia artificial, surgieron distintas clasificaciones. Una de las más reconocidas es la distinción entre IA débil (weak AI) e IA fuerte (strong AI), presentada por John Searle en 1980. La primera se refiere a sistemas que actúan como si fueran inteligentes, simulando comportamientos cognitivos sin poseer verdadera comprensión o conciencia. La segunda se emplea actualmente para describir lo que también se conoce como IA general (general AI o human-level AI): programas capaces de resolver una amplia variedad de tareas, incluso aquellas para las que no fueron entrenados, con un desempeño comparable al humano (Russell & Norvig, 2021, p. 1032).

En el sector financiero, la aplicación de la inteligencia artificial ha dejado de ser un elemento experimental para convertirse en un recurso estratégico que impulsa la eficiencia

operativa y la personalización del servicio. Los asistentes virtuales, el análisis predictivo y la detección de fraude son ejemplos de cómo la IA puede optimizar procesos críticos y, al mismo tiempo, mejorar la experiencia del cliente. Sin embargo, al ser una industria altamente regulada y la sensibilidad de los datos manejados exigen que la adopción de estas tecnologías se realice bajo criterios estrictos de seguridad, transparencia y cumplimiento normativo. En este sentido, Tavasoli et al. (2025) proponen un marco estructurado de toma de decisiones compuesto por seis elementos interconectados, esenciales para una adopción responsable de LLM en entornos financieros. Este marco abarca un ciclo de vida más amplio que el considerado habitualmente, comenzando con la evaluación preliminar de los beneficios frente a métodos más simples y continuando con el establecimiento de políticas sólidas de gobernanza de datos, la gestión integral de riesgos, la supervisión ética, la justificación del retorno de inversión (ROI) y las estrategias de implementación (p. 2).

3.1.3 Aprendizaje automático

El aprendizaje automático (machine learning) es el área de la inteligencia artificial dedicada a desarrollar algoritmos capaces de mejorar su rendimiento en una tarea a medida que se procesan más datos, sin que sea necesario programar explícitamente cada instrucción para cada hilo de ejecución. Según Burkov (2023, p. 19), este enfoque se basa en recopilar un conjunto de datos (dataset), a partir del cual se construye un modelo matemático que aprende patrones para posteriormente resolver problemas o realizar predicciones. El valor del aprendizaje automático radica en su capacidad de generalizar, es decir, de aplicar el conocimiento adquirido durante el entrenamiento a nuevos casos no vistos previamente, característica esencial para sistemas que deben operar en entornos cambiantes, como los servicios financieros digitales.

En el contexto de los modelos de lenguaje a gran escala, el aprendizaje automático es el mecanismo que permite que estas arquitecturas adquieran la capacidad de comprender y generar lenguaje humano. Al entrenarse con grandes volúmenes de datos textuales, los LLM aprenden patrones complejos que luego pueden aplicarse a tareas específicas, como responder consultas, resumir información o detectar intenciones en un diálogo.

Existen tres enfoques principales dentro del aprendizaje automático. El aprendizaje supervisado utiliza datos etiquetados para que el modelo aprenda a predecir resultados conocidos. En banca, por ejemplo, se aplica en la evaluación crediticia, donde las características de un solicitante se comparan con historiales previos para estimar el riesgo de impago. Como señalan Russell y Norvig (2021, p. 671), este enfoque se basa en entrenar el

sistema con pares entrada-salida para que pueda generalizar a nuevos casos. El aprendizaje no supervisado, en cambio, permite que el agente descubra patrones en los datos sin retroalimentación explícita. La tarea más común en este enfoque es el agrupamiento (clustering), que consiste en identificar grupos de ejemplos similares. Por ejemplo, un sistema de visión por computadora puede encontrar un gran grupo de imágenes que un hablante de inglés llamaría “gatos” a partir de imágenes previas que representan a dicho animal. Por último, el aprendizaje por refuerzo entrena al modelo mediante una serie de refuerzos (recompensas y castigos). Por ejemplo, al final de una partida de ajedrez el agente recibe la indicación de que ha ganado (recompensa) o perdido (castigo). Es responsabilidad del agente identificar qué acciones previas fueron determinantes y ajustar su comportamiento para obtener más recompensas en el futuro.

En la siguiente tabla se visualiza una clasificación de acuerdo con lo explicado por Russell y Norvig (2021):

Tabla 2

Características de los tipos de aprendizaje automático

Característica/Tipo	Aprendizaje Supervisado	Aprendizaje No Supervisado	Aprendizaje por Refuerzo
Datos	Etiquetados	Etiquetados y no etiquetados	Interacción con un entorno (contexto)
Aplicación	Predice salidas conocidas	Descubre patrones o grupos	Maximiza recompensas
Ejemplo en banca	Evaluación crediticia de un cliente	Segmentación de clientes para ofrecer productos	Gestión de carteras de clientes

Nota. Elaboración propia.

El aprendizaje automático constituye, por tanto, la base sobre la que se desarrollan los modelos de lenguaje a gran escala al aplicar arquitecturas capaces de aprender patrones complejos del lenguaje a partir de grandes volúmenes de datos. Estos modelos no solo heredan los principios de los enfoques supervisado, no supervisado y por refuerzo, sino que los combinan para optimizar su capacidad de generar, interpretar y contextualizar tareas que deberán resolver.

En el sector bancario, se aplica el aprendizaje automático en asistentes virtuales que no solo responden a preguntas frecuentes, sino que también interpretan el contexto de cada interacción y adaptan las respuestas en tiempo real. Esta integración permite construir asistentes virtuales que interactúan de manera más natural con los clientes, interpretando sus solicitudes y adaptando las respuestas a las políticas y datos de la entidad. En el siguiente apartado se abordará el funcionamiento interno de la IA generativa, clave para comprender cómo los LLM llevan a cabo estas tareas de forma autónoma y eficiente.

3.1.4 Funcionamiento de la IA Generativa

La inteligencia artificial generativa funciona a partir de modelos de aprendizaje profundo que identifican patrones en grandes volúmenes de datos y los utilizan para generar contenido nuevo, coherente y contextual. En lugar de limitarse a reglas predefinidas, estos modelos aprenden representaciones del lenguaje a través del entrenamiento sobre textos reales, lo que les permite simular la producción de lenguaje humano en tareas diversas como traducción, redacción automática o respuestas con conocimiento histórico (Bommasani et al., 2021, p. 3).

En el procesamiento de lenguaje natural (PLN), los modelos generativos como GPT operan prediciendo la siguiente palabra en una secuencia de texto a partir del contexto anterior. Este mecanismo probabilístico permite que las respuestas generadas por el sistema mantengan coherencia semántica y sintáctica. Para ello, los modelos requieren un preentrenamiento masivo sobre corpus textuales extensos, seguido por un proceso de ajuste fino (fine-tuning) que les permite adaptarse a tareas o dominios específicos (Bommasani et al., 2021, p. 10). Esta tecnología resulta especialmente útil en entornos empresariales, donde los modelos pueden comprender mejor la intención detrás de las consultas del usuario, generar respuestas más precisas y adaptarse al contexto organizacional. Según Chen et al. (2023, p. 10), las empresas que incorporan GenAI en sus procesos de atención logran no solo reducir costos operativos, sino también aumentar la personalización de los servicios y la satisfacción del cliente.

Mikolov et al. (2013, p. 1) presenta un análisis de la inteligencia artificial generativa como representación del lenguaje en forma matemática. En lugar de tratar las palabras como símbolos aislados, los modelos convierten cada término en un vector continuo en un espacio de alta dimensión denominado “embedding” (representaciones vectoriales). De esta manera, palabras con significados similares se ubican próximas en dicho espacio, lo que permite capturar relaciones semánticas y sintácticas relevantes para el procesamiento automático del lenguaje. Mikolov et al. (2013) propusieron dos arquitecturas fundamentales para el

aprendizaje de estas representaciones: “Continuous Bag-of-Words (CBOW)” y “Skip-gram”. La primera predice la palabra central a partir de las que la rodean, mientras que la segunda toma una palabra y busca anticipar su contexto cercano (p. 4).

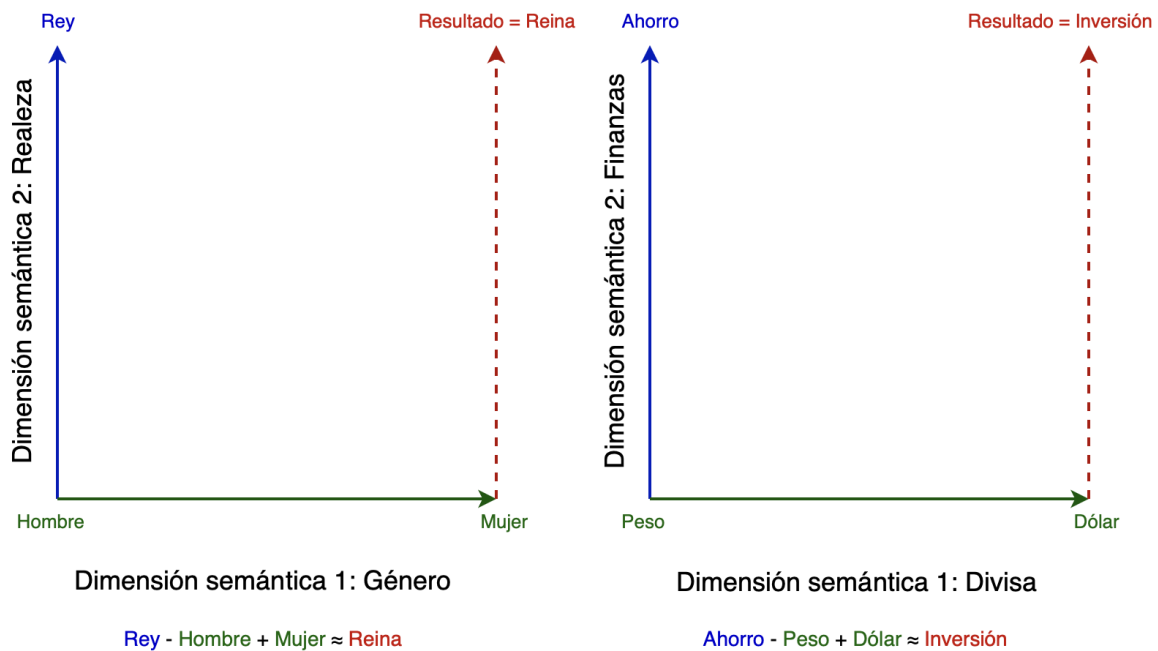
El impacto de estas representaciones fue decisivo porque demostraron que, al escalar el entrenamiento a miles de millones de palabras, los vectores podían capturar relaciones lingüísticas de manera lineal. Mikolov et al. (2013) mostraron que estas representaciones permiten identificar no solo palabras similares en sentido general, sino también múltiples grados de similitud, como sufijos compartidos o variaciones morfológicas. La similitud entre vectores trasciende las regularidades sintácticas básicas y posibilita el descubrimiento de analogías semánticas complejas. Un ejemplo clásico es la operación vectorial $\text{vector}(\text{“Rey”}) - \text{vector}(\text{“Hombre”}) + \text{vector}(\text{“Mujer”})$, que resulta cercana al vector de la palabra Reina (p. 3). En el espacio vectorial, las diferencias entre palabras representan relaciones semánticas. Por esa razón la operación Rey – Hombre captura la noción de realeza masculina, mientras que + Mujer añade la dimensión de género femenino. El resultado se aproxima entonces a Reina, mostrando cómo los embeddings permiten resolver analogías mediante operaciones algebraicas simples (Mikolov et al., 2013, p. 2).

La capacidad de los *embeddings*, representaciones vectoriales, para capturar regularidades lingüísticas puede trasladarse también a dominios específicos como el financiero. Al entrenar vectores en grandes corpus de datos bancarios, es posible identificar relaciones semánticas propias del sector. Esto aplicaría entonces también a una operación vectorial hipotética como $\text{vector}(\text{“Ahorro”}) - \text{vector}(\text{“Peso”}) + \text{vector}(\text{“Dólar”})$ que podría acercarse al vector de la palabra “Inversión”, reflejando la asociación entre el ahorro en moneda local, la referencia a una divisa extranjera y el concepto financiero resultante. Este tipo de relaciones permiten que los modelos generativos comprendan mejor el contexto de las consultas y ofrezcan respuestas con mayor precisión y coherencia al dominio específico.

En la representación vectorial del lenguaje, los embeddings permiten que las relaciones semánticas y sintácticas entre palabras se traduzcan en operaciones matemáticas dentro de un espacio multidimensional (Mikolov et al., 2013, p. 2). La Figura 7 ilustra este principio con dos ejemplos: una analogía clásica de tipo lingüístico y una analogía adaptada al contexto financiero.

Figura 7

Representación vectorial de relaciones lingüísticas



Nota. Elaboración propia.

La Figura 7 representa en el primer caso, la operación $\text{Rey} - \text{Hombre} + \text{Mujer} \approx \text{Reina}$ evidencia cómo el modelo identifica la relación de realeza y la proyecta hacia el género femenino, resultando en el concepto de Reina. De manera análoga, en el ejemplo financiero se observa la operación $\text{Ahorro} - \text{Peso} + \text{Dólar} \approx \text{Inversión}$, donde la noción de valor financiero se combina con el cambio de divisa, aproximando el resultado al concepto de inversión. Las flechas representan las dimensiones semánticas involucradas, y la posición de los puntos demuestra que las diferencias entre vectores codifican relaciones conceptuales que los modelos pueden manipular algebraicamente para inferir significados.

Las operaciones algebraicas en el espacio vectorial permiten resolver analogías de distinta naturaleza y no se limitan a un único caso. Como señalan Mikolov et al. (2013), los vectores de palabras con este tipo de relaciones semánticas pueden utilizarse para mejorar procesos actuales como la traducción automática, la recuperación de información y los sistemas de respuesta a preguntas. Según Mikolov et al. (2013), estas representaciones capturan regularidades gramaticales y semánticas, por ejemplo, la formación de plurales (ratón-ratones; dólar-dólares), tiempos verbales (caminar-caminó o nadar-nadó) y pares país-capital (Francia-París; Alemania-Berlín). Estas evidencias sustentan que los embeddings representan reglas

lingüísticas implícitas mediante operaciones matemáticas simples, habilitando su aplicación en múltiples tareas de PLN (p. 5).

3.1.5 Arquitecturas de Modelos Generativos

Los primeros intentos de modelar secuencias lingüísticas mediante redes neuronales se apoyaron en el trabajo de Elman (1990), quien introdujo la idea de utilizar conexiones recurrentes para dotar a las redes de una memoria dinámica. Este mecanismo permitía que los estados internos de la red se retroalimentaran, de modo que las representaciones desarrolladas reflejaran no solo la entrada actual, sino también el contexto de estados previos. Según Elman (1990, p. 179), se utilizan conexiones recurrentes con el fin de dotar a las redes de una memoria dinámica, logrando así que los patrones de las unidades ocultas se retroalimentan a sí mismos. El autor realizó un procedimiento para entrenar Redes Neuronales Recurrentes (RNR) con oraciones artificiales que incluían categorías específicas de sustantivos y verbos, utilizando un programa generador de oraciones para construir un conjunto de enunciados cortos de dos y tres palabras. Se eligieron trece clases de sustantivos y verbos. Es decir que el experimento partía de categorías gramaticales explícitas que la red debía procesar, y las representaciones ocultas resultantes lograban capturar esas diferencias (p. 195).

Una siguiente evolución de arquitecturas para el modelo del lenguaje natural lo realizaron Bengio et al. (2003), quienes propusieron un modelo probabilístico neuronal del lenguaje. A diferencia de los modelos tradicionales, que generalizaban únicamente a partir de combinaciones cortas y frecuentes de palabras, este enfoque de Bengio et al. (p. 1138) propone combatir la dimensionalidad aprendiendo una representación distribuida de las palabras que permita que cada oración de entrenamiento informe al modelo acerca de un número exponencial de oraciones vecinas semánticamente relacionadas. El principal aporte de este planteamiento fue la introducción de vectores de características continuas para cada palabra, de modo que términos con roles semánticos o gramaticales similares quedaran representados en regiones próximas del espacio. De este modo, el modelo era capaz de asignar alta probabilidad a secuencias nunca vistas, siempre que sus palabras compartieran propiedades con ejemplos previos. En el modelo propuesto, se generaliza de esta manera porque se espera que las palabras similares tengan un vector de características también similar, y dado que la función de probabilidad es una función continua de estos valores de características, un pequeño cambio en las características inducirá un pequeño cambio en la probabilidad. Por lo tanto, la presencia de solo una de las oraciones anteriores en los datos de entrenamiento aumentará la probabilidad, no solo de esa oración, sino también de su número combinatorio de palabras vecinas en el

espacio de oraciones (representadas por secuencias de vectores de características) (Bengio et al., 2003, p. 1140).

Un avance decisivo para superar las limitaciones de las RNN que expuso Elman (1990), fue la introducción de las redes “Long Short-Term Memory (LSTM)” de Hochreiter y Schmidhuber (1997). Estas redes resolvieron el problema de la desaparición del gradiente al incorporar un mecanismo de celdas de memoria con compuertas multiplicativas, capaces de decidir qué información conservar, actualizar u olvidar en cada paso. Según Hochreiter y Schmidhuber (1997), este diseño permitió que las LSTM gestionaran dependencias de largo alcance en secuencias, algo crítico en tareas de procesamiento de lenguaje natural (p. 1).

3.1.6 Agentes Autónomos y Arquitecturas Híbridas en IA

Las arquitecturas basadas en Transformers, explicadas anteriormente por Vaswani, constituyen el núcleo técnico de los agentes autónomos modernos, ya que su capacidad de autoatención y de procesamiento paralelo facilita la integración de memoria, planificación y razonamiento en contextos dinámicos. Estas propiedades permiten que los modelos de lenguaje a gran escala no solo generen texto coherente, sino que se conviertan en componentes capaces de interactuar de manera continua con entornos reales, adaptándose a las demandas del usuario y ejecutando acciones proactivas. En este sentido, Russell y Norvig (2021) definen a un agente inteligente como algo que percibe su entorno a través de sensores y actúa sobre ese entorno a través de actuadores (p. 54). Esta concepción permite entender a los agentes autónomos potenciados por LLM como sistemas que no solo producen lenguaje, sino que también perciben información de múltiples fuentes como bases de datos, registros de clientes, datos provenientes de otros sistemas a través de integraciones, y que luego actúan en consecuencia para alcanzar objetivos definidos.

3.2 Procesamiento de Lenguaje Natural

3.2.1 Introducción

El procesamiento de lenguaje natural constituye un área de la inteligencia artificial dedicada al diseño de algoritmos capaces de interpretar y generar lenguaje humano. Según Eisenstein (2019), el PLN se centra en el diseño y análisis de algoritmos y representaciones computacionales para procesar el lenguaje humano natural, por lo que tiene como objetivo aportar nuevas capacidades computacionales en torno al lenguaje humano. Esto puede ser representado en actividades como extraer información de textos, traducir entre lenguas, responder preguntas, mantener una conversación o seguir instrucciones, entre otras

aplicaciones (p. 14). En la misma línea, Goldberg (2017) define al PLN como un término colectivo que hace referencia al procesamiento computacional automático de lenguas humanas. Es decir que incluye a los algoritmos que reciben como entrada texto producido por humanos, y también a los algoritmos que producen texto con apariencia natural como salida (p. 20). Ambas perspectivas coinciden en destacar que el PLN no se limita a la traducción o clasificación textual, sino que constituye la base de las interacciones entre personas y sistemas inteligentes, fundamento indispensable para el diseño de asistentes virtuales conversacionales.

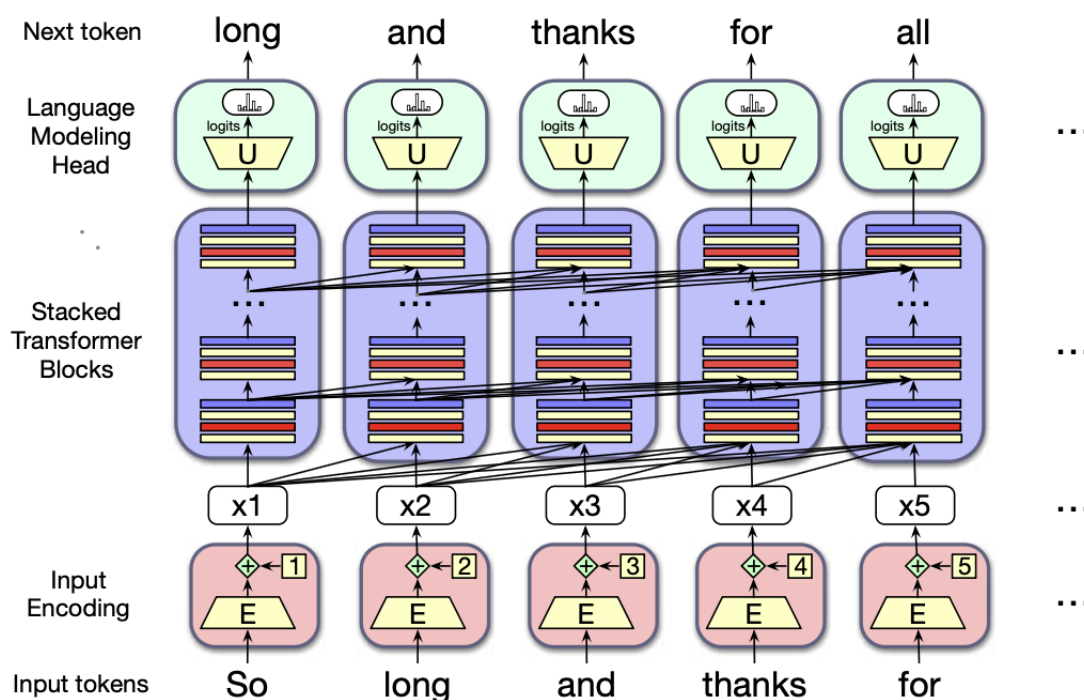
La evolución del PLN refleja el tránsito desde los enfoques simbólicos basados en reglas hacia modelos estadísticos y más recientemente hacia arquitecturas neuronales profundas. Manning y Schütze (1999) subrayan que los métodos estadísticos de PLN han liderado el camino en proveer desambiguación exitosa en sistemas de gran escala que procesan texto real, lo que permitió superar las limitaciones de las gramáticas manuales en entornos complejos (p. 19). Con la irrupción de las redes neuronales, el campo experimentó un cambio decisivo. Goldberg (2017) recuerda que los primeros intentos fueron simbólicos porque el lenguaje natural es altamente ambiguo y variable, lo que condujo a la ciencia hacia la adopción de un enfoque estadístico. El autor luego comenta que alrededor de 2014, el campo comenzó a ver cierto éxito al cambiar de estos modelos lineales con entradas dispersas a modelos de redes neuronales no lineales con entradas densas, y complementa que los avances condujeron a progresos en acciones como el modelado del lenguaje y la traducción automática (p. 20).

La consolidación del PLN en la última década estuvo marcada por la irrupción de los Transformers en conjunto con los LLM. Según Jurafsky y Martin (2023), los Transformers se aplicaron inicialmente como “codificadores-decodificadores” y los modelos de lenguaje de gran escala basados en Transformers han cambiado por completo el campo del procesamiento del habla y el lenguaje (p. 170). Los autores agregan en este sentido que un transformer puede ser entendido como una red neuronal con una estructura específica que incluye un mecanismo de atención multi-cabecal, es decir, como una forma de construir representaciones contextuales del significado de un token, atendiendo e integrando información de los tokens circundantes, lo que ayuda al modelo a aprender cómo los tokens se relacionan entre sí a lo largo de grandes secuencias (p. 171).

La arquitectura de un Transformer representa diferentes capas. Jurafsky y Martin (2023) presentan una aproximación gráfica en la siguiente imagen:

Figura 8

Arquitectura de un transformador



Nota. Obtenido de *The Architecture of (left-to-right) transformer* (p. 171), por Jurafsky y Martin, 2023, *Speech and Language Processing*.

La Figura 8 de Jurafsky y Martin (2023) presenta que cada token de entrada se debe codificar, luego pasar a través de un conjunto de bloques apilados que forman el transformador y posteriormente por un modelo de lenguaje que predice el siguiente token. En la imagen, el primer token está representado por la palabra “long” y luego el segundo por “and”, como una referencia de la consecuencia de tokens que ingresan al transformador (p. 171).

El modelo de Transformers se organiza en una arquitectura compuesta por tres elementos principales. En su núcleo se ubican columnas de bloques de transformer, cada uno de los cuales es una red multicapa integrada por mecanismos de atención multi-cabezal, redes *feed-forward* (red de avance) y pasos de normalización. Estos bloques transforman los vectores de entrada correspondientes a cada token en vectores de salida contextualizados, mapeando así ventanas completas de secuencias de tokens a representaciones de igual longitud. La profundidad de la red puede variar desde una docena hasta más de noventa bloques apilados. Previo a estos bloques, el modelo cuenta con un componente de codificación de entrada, que convierte cada token en un vector contextual mediante una matriz de embeddings y un sistema de codificación posicional. Una vez procesados, los vectores pasan al componente final, la

cabeza de modelado del lenguaje, que utiliza una matriz de desempaquetado y una función *softmax* (función de normalización) sobre el vocabulario para generar la predicción del siguiente token. De este modo, el transformer articula la codificación de entradas, la autoatención multicabecal y la generación de tokens como pilares de su funcionamiento (Jurafsky & Martin, 2023, p. 172).

3.2.2 Técnicas y Enfoques Principales en PLN

Los primeros sistemas de procesamiento de lenguaje natural se desarrollaron bajo un paradigma basado en reglas, en el que el conocimiento lingüístico se codificaba de manera explícita mediante gramáticas, diccionarios y sistemas lógicos. Estos enfoques lograron avances iniciales en tareas como la traducción automática y el análisis sintáctico, pero enfrentaban algunas dificultades para obtener buenas respuestas como producto de la ambigüedad y la variabilidad del lenguaje humano. Según Jurafsky y Martin (2023), los sistemas simbólicos dependían de reglas escritas manualmente y resultaban poco escalables frente a la complejidad del lenguaje humano (p. 375). Estas limitaciones motivaron la transición hacia modelos estadísticos, que proponían aprender las regularidades lingüísticas a partir de datos en lugar de reglas predefinidas. Con la llegada de los enfoques estadísticos, el procesamiento de lenguaje natural dio un salto metodológico al pasar de reglas manuales a modelos entrenados con datos reales.

Puntualmente, el enfoque estadístico en PLN se consolidó con el uso de los modelos *n*-gram, que permitían estimar la probabilidad de una palabra a partir de un contexto limitado. En lugar de calcular la probabilidad de un término dado todo el historial previo de palabras (una tarea inabordable con corpus finitos) se recurrió a la suposición de Markov, que aproxima el contexto considerando solo un número reducido de palabras anteriores. Según Jurafsky y Martin (2023), la intuición del modelo *n*-gram consiste en que, en lugar de calcular la probabilidad de una palabra dada toda su historia, es posible aproximar esa historia usando solo las últimas palabras, lo que se conoce como la suposición de Markov (p. 39). Esta simplificación permitió aplicar métodos probabilísticos de manera práctica en tareas como el modelado del lenguaje y la predicción de palabras.

Un caso paradigmático de la aplicación de modelos estadísticos en PLN fue la traducción automática estadística (SMT, por sus siglas en inglés). La propuesta consistía en estimar, a partir de grandes corpus paralelos, la probabilidad de que una oración en un idioma correspondiera a otra en la lengua de destino. Koehn (2010) plantea en su investigación una problemática en la traducción automática estadística porque se centraba en la traducción de

oraciones individuales. Si bien la traducción de un idioma a otro casi siempre es posible traduciendo textos oración por oración (con la ocasional división y fusión de oraciones adyacentes), algunos problemas de traducción no pueden resolverse únicamente a nivel de oración. En este sentido, según Koehn (2010) uno de los principales problemas es la traducción de correferencias o anáforas. Si bien la primera mención de una entidad suele estar detallada (por ejemplo, el presidente George Bush), posteriormente puede mencionarse solo con un pronombre (él) o una descripción abreviada (el presidente). Esto puede generar ambigüedad en oraciones posteriores. Basta solo con imaginar la complejidad si la palabra “presidente” se necesita de forma diferente para un cargo público masculino o femenino, o incluso si debe ser diferenciada para el tipo de institución como una empresa, un club estudiantil o incluso un país. Es por ello por lo que Koehn asegura que una mención posterior puede no tener suficiente información, por lo que es necesario retroceder a las menciones anteriores, un problema conocido como resolución de anáforas. (Koehn, 2010, p. 52). Este enfoque se implementaba combinando dos componentes principales: un modelo de traducción, encargado de establecer las correspondencias probabilísticas entre secuencias de palabras en distintos idiomas, y un modelo de lenguaje, cuya función era garantizar que la salida resultara gramaticalmente correcta y coherente en el idioma de destino. Gracias a esta arquitectura, la SMT dominó el campo de la traducción automática por más de una década, hasta ser desplazada por los modelos neuronales. (Koehn, 2010, p. 7).

A pesar de sus avances, la SMT presentaba limitaciones significativas. En primer lugar, su rendimiento dependía fuertemente de la disponibilidad de grandes corpus paralelos, un recurso escaso en la mayoría de los pares de lenguas (p. 6). Además, los modelos estadísticos enfrentaban problemas derivados de la escasez de datos: al aumentar la complejidad del modelo, el número de parámetros crecía de forma exponencial, y aun con corpus extensos resultaba insuficiente para cubrir todos los contextos posibles. De acuerdo con Manning y Schütze (1999), incluso si se logra dividir los datos en demasiados grupos, habrá muchos parámetros que estimar (p. 193).

3.2.3 Rol del PLN en Asistentes Virtuales

Los asistentes virtuales que funcionan con el procesamiento de lenguaje natural tienen una gran ventaja contra otros tipos de asistentes predictivos, porque pueden adaptar al contexto en el que funcionan. Según Goldberg (2017), las redes neuronales proporcionan una potente maquinaria de aprendizaje muy atractiva para su uso en problemas de lenguaje natural. Esto ocurre gracias a la aplicación del aprendizaje profundo, entendido como una rama del

aprendizaje automático, de las redes neuronales, compuesta por un conjunto de técnicas de aprendizaje inspiradas históricamente en el funcionamiento de la computación en el cerebro y que se caracterizan por el aprendizaje de funciones matemáticas diferenciables parametrizadas (p. 28).

Goldberg también detalla el gran beneficio que esto considera para la comprensión del contexto y la generación de respuestas que pueden acercarse a la forma en la que los seres humanos se comunican porque si bien todo aprendizaje automático se caracteriza por aprender a realizar predicciones basadas en observaciones previas, los enfoques de aprendizaje profundo funcionan aprendiendo no solo a predecir, sino también a representar correctamente los datos, de modo que sean adecuados para la predicción. Dado un amplio conjunto de asignaciones de entrada-salida deseadas, los enfoques de aprendizaje profundo funcionan introduciendo los datos en una red que produce transformaciones sucesivas de los datos de entrada hasta que una transformación final predice la salida. Las transformaciones producidas por la red se aprenden a partir de las asignaciones de entrada-salida dadas, de modo que cada transformación facilita la relación de los datos con la etiqueta deseada. (Goldberg, 2017, p. 28).

Los sistemas modernos no solo interpretan intenciones y extraen entidades, sino que además incorporan mecanismos de aprendizaje continuo para mejorar con cada interacción. Según McTear (2020), los sistemas de diálogo actuales han superado muchas de las limitaciones porque tienen la capacidad de aprender de la experiencia, a diferencia de los sistemas anteriores que eran estáticos y no alteraron ni mejoraron su comportamiento con el tiempo (p. 29). Esto refuerza que el PLN no solo permite comprender la entrada del usuario, sino también mantener diálogos adaptativos que elevan la experiencia de interacción.

3.3 Modelado de Lenguaje a Gran Escala

3.3.1 Introducción

Los modelos de lenguaje a gran escala se han convertido en el núcleo de la revolución actual en el procesamiento de lenguaje natural. Según Jurafsky y Martin (2023), su principal característica radica en la capacidad de aprender representaciones del lenguaje mediante un entrenamiento basado en la predicción de la siguiente palabra en grandes corpus de texto. Este procedimiento repetitivo, aplicado a volúmenes masivos de datos, permite que el modelo adquiera conocimiento no solo lingüístico, sino también contextual y factual, lo que amplía su capacidad de generalización frente a tareas variadas en entornos reales. Según los autores, el

modelado de lenguaje es un sistema computacional que puede predecir una siguiente palabra a partir de palabras previas (Jurafsky & Martin, 2023, p. 146).

El aumento en la escala es otro de los factores distintivos. Zhao et al. (2023) destacan que los LLM modernos suelen estar contruidos sobre la arquitectura Transformer y pueden alcanzar cientos de miles de millones de parámetros, entrenados con colecciones textuales de magnitud sin precedentes. La combinación de tamaño y datos posibilita que los modelos superen el rendimiento de métodos anteriores, ya que capturan dependencias a largo plazo en el lenguaje y manejan variaciones complejas que antes resultaban difíciles de modelar con enfoques estadísticos o neuronales más simples.

Otro aspecto relevante es la aparición de capacidades emergentes, entendidas como habilidades que no fueron programadas explícitamente, pero que surgen a partir de la escala alcanzada por los modelos. Tanto Jurafsky y Martin (2023) como Zhao et al. (2023) coinciden en que esta propiedad permite a los LLM adaptarse a una gran diversidad de tareas, desde el resumen automático de documentos hasta la resolución de preguntas complejas, pasando por la generación de respuestas en sistemas conversacionales. Estas características los han consolidado como la tecnología fundamental para asistentes virtuales avanzados, que requieren no solo comprender consultas, sino también contextualizar y producir interacciones coherentes en tiempo real.

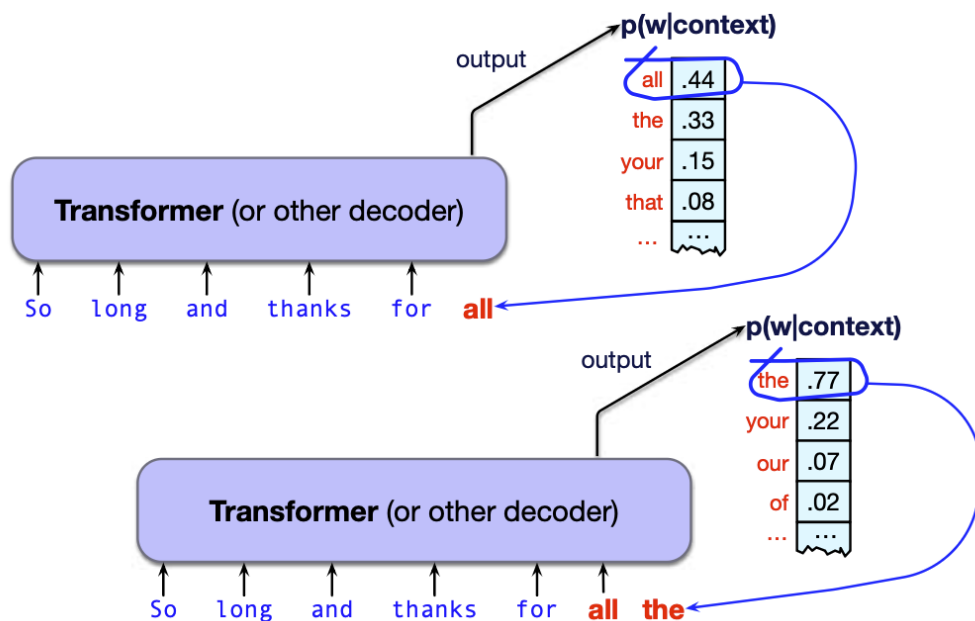
3.3.2 Proceso de Preentrenamiento y Ajuste Fino

El preentrenamiento constituye la base de los modelos de lenguaje a gran escala. Jurafsky y Martin (2023) explican que estos modelos aprenden de manera auto-supervisada a partir de la tarea de predecir la siguiente palabra en un corpus de gran tamaño, lo que les permite adquirir conocimiento lingüístico, contextual y factual sin intervención manual. El procedimiento repetitivo de entrenamiento convierte al LLM en un modelo generalista, capaz de transferir lo aprendido a múltiples aplicaciones, desde la traducción automática hasta la interacción conversacional.

La Figura 9 presenta la conformación de los modelos autorregresivos de izquierda a derecha, en los cuales cada token producido pasa a formar parte del contexto y se utiliza como insumo para determinar la predicción del siguiente elemento.

Figura 9

Modelo predictivo con tokens



Nota. Obtenido de *Predictive Model* (p. 141), por Jurafsky y Martin, 2023, *Speech and Language Processing*

Jurafsky y Martin (2003) explican que, de las arquitecturas empleadas en el modelado de lenguaje, los modelos tipo decoder son los más utilizados en los LLM modernos. Estos sistemas reciben como entrada una secuencia de tokens y generan la salida de manera autoregresiva, es decir, prediciendo un token a la vez de izquierda a derecha, en función únicamente de los tokens previos en la secuencia. Al ser modelos generativos, tienen la capacidad de producir nuevo texto, lo que los hace adecuados para tareas de generación libre como redacción, asistencia conversacional o resolución de preguntas. Ejemplos de arquitecturas de este tipo incluyen GPT, Claude, LLaMA y Mistral, que representan la base de los asistentes virtuales actuales. En contraste, los modelos tipo *encoder* (codificador) procesan una secuencia de tokens y producen para cada elemento una representación vectorial contextualizada, considerando tanto los elementos previos como posteriores en la secuencia. Estos modelos suelen entrenarse bajo el paradigma de *masked language modeling*, en el que una palabra se enmascara y la red aprende a predecirla a partir del contexto circundante. A diferencia de los *decoders* (decodificadores), los *encoders* no son generativos; su principal aplicación se encuentra en tareas de clasificación supervisada, como análisis de sentimiento, categorización temática o extracción de entidades. Dentro de este grupo destacan modelos

como BERT y RoBERTa, ampliamente utilizados como base para sistemas de comprensión de lenguaje. Por último, la arquitectura encoder-decoder combina ambas estrategias. En este caso, el encoder procesa la secuencia de entrada y genera una representación intermedia, mientras que el decoder produce una secuencia de salida condicionada por esa representación. Lo distintivo de este esquema es que no existe una relación estricta entre la longitud o el tipo de tokens de entrada y los de salida: pueden pertenecer a conjuntos diferentes e incluso tener longitudes muy distintas. Este diseño resulta especialmente útil para tareas de mapeo entre modalidades o lenguajes distintos, como la traducción automática, en los que la entrada está en un idioma y la salida en otro, o el reconocimiento de voz, en el cual la entrada corresponde a señales de audio y la salida a tokens de texto. Modelos como T5 y BART se enmarcan en esta categoría.

La magnitud de los datos y parámetros implicados exige estrategias computacionales avanzadas. Zhao et al. (2023) señalan que entrenar LLM modernos requiere técnicas de paralelismo (data, tensor, pipeline), algoritmos de reducción de memoria como ZeRO y el uso de precisión mixta, lo que hace posible escalar a centenares de GPU en funcionamiento. Estos avances en infraestructura han permitido que los modelos base capturen regularidades lingüísticas que luego pueden adaptarse a dominios específicos, como el financiero.

Una vez entrenado el modelo base, es necesario alinearlo y adaptarlo. Zhao et al. (2023) describen dos etapas centrales de post-training: el ajuste fino por instrucciones (instruction tuning), que refina el modelo con datos de instrucciones y respuestas, y la alineación con preferencias humanas (human preference-align), que incluye métodos como RLHF para ajustar el comportamiento del modelo según la retroalimentación humana. De este modo, se mejora la capacidad del sistema para seguir instrucciones, evitar sesgos y producir respuestas más útiles.

En el caso bancario, la aplicación de estas técnicas es clave: un asistente virtual puede aprovechar un LLM generalista, luego ajustarse con datos de políticas internas y manuales de atención al cliente mediante ajuste fino por instrucciones, y finalmente refinarse con corpus financieros mediante PEFT o ajuste de dominio. Este *pipeline* (flujo de procesamiento) asegura un modelo que no solo entiende consultas generales, sino que también responde con precisión en el marco regulatorio y terminológico específico del sector.

3.3.3 Ingeniería de Prompts y Optimización de Resultados

Según IBM (s.f.b), la ingeniería de prompts “ayuda a moldear el resultado del modelo, lo que garantiza que la inteligencia artificial responda de manera significativa y coherente”. Es decir que es un aspecto clave para generar instrucciones que los modelos de inteligencia

artificial generativa utilizan para “comprender y responder mejor a una amplia gama de consultas” (IBM, s. f.b).

Golani (2025) define la ingeniería de prompts como el diseño, refinamiento y optimización de las instrucciones que se entregan a un modelo de lenguaje, con el fin de orientar sus respuestas sin necesidad de modificar su arquitectura interna (p. 5). Este enfoque se diferencia del ajuste fino, ya que no requiere reentrenar el modelo ni disponer de grandes volúmenes de datos, sino que aprovecha la capacidad latente del modelo base.

Entre sus principales beneficios se destacan la reducción de costos, al eliminar la necesidad de entrenar modelos completos y permitir el uso de servicios a través de interfaces de programación (API) con un esquema de pago por uso. Además, ofrece una alta agilidad y capacidad de iteración, dado que los prompts pueden modificarse y evaluarse en lapsos cortos, lo que habilita ciclos de prueba y error mucho más rápidos que los procesos de reentrenamiento. A ello se suma la baja barrera de entrada, ya que no requiere conocimientos técnicos avanzados: especialistas en negocio o diseñadores pueden participar en la creación de prompts, favoreciendo un desarrollo interdisciplinario y colaborativo (Golani, 2025, p. 6). Otro aspecto positivo es la flexibilidad, ya que los prompts diseñados para un modelo base suelen adaptarse de manera inmediata a las mejoras que introducen nuevas versiones, lo que asegura un proceso de actualización prácticamente automático (Golani, 2025, p. 7).

No obstante, el autor indica que la ingeniería de prompts también presenta limitaciones. Una de ellas es su fragilidad, dado que pequeñas modificaciones en la redacción pueden ocasionar pérdidas significativas en el rendimiento del modelo. Asimismo, el uso de prompts extensos genera un alto consumo de tokens, lo que reduce el espacio disponible para entradas del usuario y encarece el costo de operación (Golani, 2025, p. 7). Otro aspecto crítico es la dependencia de proveedores externos, puesto que las organizaciones que adoptan este enfoque quedan sujetas a cambios en precios, condiciones de uso o riesgos asociados al tratamiento de datos sensibles mediante servicios de terceros (Golani, 2025, p. 8).

3.3.4 Evaluación de Desempeño en LLM: Métricas

La evaluación del desempeño de los modelos de lenguaje a gran escala constituye un aspecto crítico para garantizar su utilidad en entornos de aplicación como la atención al cliente bancario. Es decir que no resulta suficiente medir únicamente la precisión de las respuestas, sino que es necesario considerar dimensiones múltiples que contemplen la coherencia, la robustez, la equidad y la eficiencia del sistema. Liang et al. (2023) plantean el marco HELM (Holistic Evaluation of Language Models), que introduce una evaluación integral basada en

siete ejes: precisión, calibración, robustez, equidad, sesgo, toxicidad y eficiencia. Esta aproximación demuestra el rendimiento de los modelos en escenarios reales, donde la calidad de la interacción no puede reducirse a una métrica única (p. 1).

La atención bancaria con agentes virtuales demanda modelos capaces de responder de forma consistente con información documental actualizada. Este requisito ha impulsado el uso de arquitecturas de Retrieval-Augmented Generation (RAG), en las cuales el modelo no se limita a generar respuestas a partir de su entrenamiento previo, sino que combina dos módulos principales: uno de recuperación de documentos relevantes desde un repositorio (por ejemplo, políticas internas, normativas bancarias o bases de conocimiento), y otro de generación de lenguaje natural, que utiliza tanto la consulta del usuario como los fragmentos recuperados para producir la respuesta final. De esta manera, RAG incrementa la capacidad del sistema para ofrecer información actualizada y reducir el riesgo de alucinaciones, ya que la salida queda anclada a evidencia verificable. De acuerdo con Oche et al. (2025), la arquitectura Retrieval-Augmented Generation (RAG) constituye un enfoque que combina técnicas de recuperación de información y generación de lenguaje natural para mejorar el resultado factual y la pertinencia de los modelos de lenguaje. RAG se implementa en tres fases: primero, la indexación de un corpus documental en un espacio vectorial de embeddings, vectores de tamaño fijo que representan fragmentos de texto, que pueden ser palabras, frases o documentos completos, y que codifican relaciones semánticas, de modo que su distancia refleja la relevancia entre consulta y pasaje; segundo, la recuperación de fragmentos relevantes según su similitud con la consulta; y tercero, la generación condicionada, en la cual el modelo de lenguaje produce la salida integrando la pregunta del usuario y los pasajes recuperados (p. 3).

La escalabilidad de los sistemas RAG ha motivado propuestas que integran infraestructuras de cómputo de alto rendimiento (HPC). En este sentido, Gökdemir et al. (2025) desarrollaron HiPerRAG, un flujo de trabajo diseñado para indexar más de 3,6 millones de artículos científicos. La arquitectura incluye el parser Oreo, capaz de extraer información multimodal de documentos en formato PDF con una eficiencia 4,5 veces superior a otros métodos, y el codificador ColTrast, un encoder sensible a la consulta que combina contrastive learning y late interaction para mejorar la precisión de la recuperación a gran escala (p. 2).

Los resultados empíricos de HiPerRAG demuestran el potencial de esta aproximación: el sistema alcanzó un 90% de exactitud en el benchmark SciQ y un 76% en PubMedQA, superando tanto a modelos entrenados específicamente en biomedicina (PubMedGPT) como a modelos generales de última generación (GPT-4). Estas métricas evidencian que la

combinación de recuperación especializada y optimización de embeddings puede aumentar la precisión y relevancia de las respuestas generadas, incluso en dominios altamente complejos (Gökdemir et al., 2025, p. 1).

Para evaluar este tipo de arquitecturas, se han desarrollado métricas específicas. El marco RAGAS propone tres dimensiones principales: fidelidad, entendida como la consistencia entre la respuesta y la evidencia recuperada; relevancia de la respuesta, que mide la alineación con la pregunta formulada; y relevancia del contexto, que evalúa la precisión y pertinencia de los fragmentos documentales seleccionados. La ventaja de este enfoque es que permite realizar evaluaciones automáticas de sistemas que integran recuperación y generación, incluso sin necesidad de disponer de respuestas de referencia anotadas por humanos (Es et al., 2024, p. 151).

3.4 Centros de Atención al Cliente

3.4.1 Inteligencia Artificial Multicanal

Como lo expresa Mashaabi et al. (2022), la atención al cliente puede definirse como el conjunto de actividades y procesos mediante los cuales una organización acompaña al consumidor antes, durante y después de la adquisición de un producto o servicio. Esta función ha evolucionado hacia un modelo multicanal en el que convergen interacciones presenciales, telefónicas y digitales, lo que incrementa la complejidad de su gestión y exige mayor coherencia y consistencia en las respuestas (p. 2). Según Mashaabi et al. (2022), el uso del PLN en la atención al cliente con aplicaciones de inteligencia artificial brinda la oportunidad de que los usuarios se comuniquen directamente con modelos de IA que pueden comprender diferentes idiomas en formatos como texto o voz, para ofrecer respuestas a los usuarios de acuerdo con las necesidades que soliciten (p. 1).

En el contexto de los servicios financieros, la atención al cliente tiene una función estratégica debido a que la confianza y la seguridad constituyen activos intangibles de alto valor. Fernández (2019) sostiene que la inteligencia artificial puede aplicarse en toda la cadena de valor del sector. De esta manera se puede incluir en tareas de procesamiento (conocido como back-office), procesos que determinan la interacción directa con los clientes (front-office), y actividades de control y monitoreo (middle-office) (p. 3). El sector financiero ha venido incorporando diferentes aplicaciones como chatbots, aplicaciones de personalización de productos para clientes, prevención de fraude, calificación crediticia, cumplimiento regulatorio y hasta sistemas biométricos de autenticación (p. 4). Estos avances permiten repensar la

experiencia de usuario (UX) no solo como un canal de soporte, sino como una experiencia integral que combina eficacia operativa con personalización y seguridad.

3.4.2 Resolución de Problemáticas Comunes

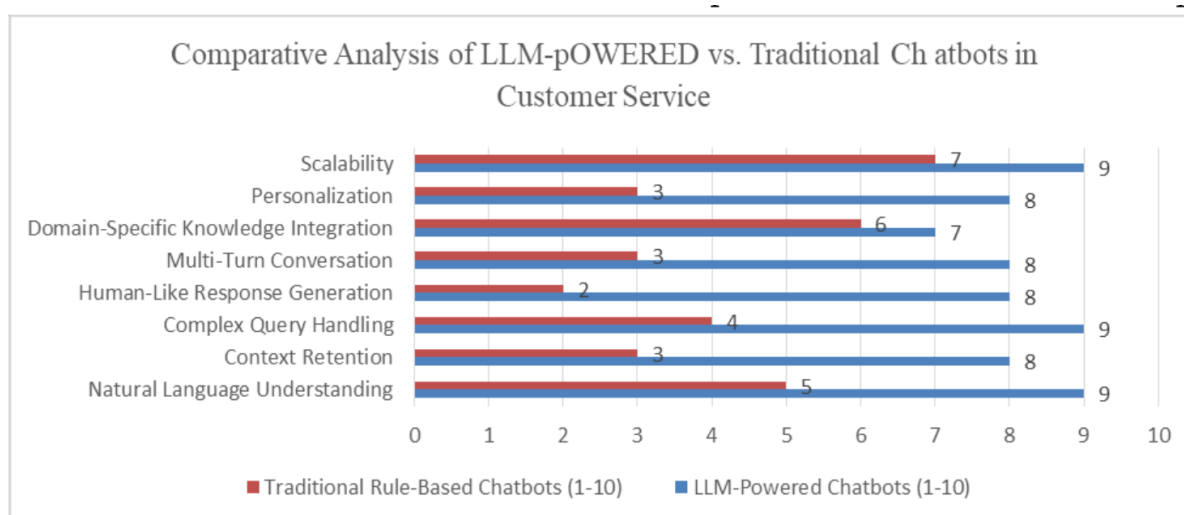
Los centros de contacto con clientes enfrentan de forma recurrente problemáticas relacionadas con la sobrecarga de consultas, la fragmentación de canales y la inconsistencia de respuestas. De acuerdo con Mashaabi et al. (2022), cerca del 75% de los clientes reportan haber experimentado servicios deficientes, lo que demuestra la necesidad de mejorar tiempos de respuesta, integración multicanal y resolución efectiva en el primer contacto (p. 2).

La automatización de tareas repetitivas mediante chatbots y asistentes virtuales se presenta como una estrategia eficaz para resolver estas dificultades. En la práctica, permiten ofrecer disponibilidad 24/7 y capturar datos de las interacciones de forma estructurada, lo que posibilita la retroalimentación continua y la mejora del servicio. Fernández (2019) señala que, en el sector bancario, la implementación de IA en la atención al cliente no solo incrementa la eficiencia operativa, sino que también habilita nuevas capacidades de análisis y control, con beneficios claros en términos de accesibilidad y adopción digital (p. 2).

Los modelos de lenguaje a gran escala (LLMs) representan un salto cualitativo en este campo. Meduri (2024) demuestra que, a diferencia de los chatbots basados en reglas, los LLMs ofrecen comprensión avanzada del lenguaje natural, mantienen contexto en conversaciones extensas y generan respuestas coherentes y personalizadas en tiempo real (p. 2). Estas características han permitido mejoras significativas en sectores como el comercio electrónico, las telecomunicaciones y los servicios financieros, donde un caso reportado evidenció un aumento del 25% en la adopción de servicios digitales gracias a la integración de controles regulatorios en un chatbot potenciado por LLM (p. 5). Según el autor, en el contexto de la atención al cliente, los lenguajes a gran escala que se comienzan a implementar en las empresas han logrado transformar a los tradicionales chatbots que se construían con reglas básicas en agentes conversacionales sofisticados (p. 2). Meduri (2024) presenta en la siguiente imagen una descripción de las principales habilidades en las que los LLM han superado a los chatbots tradicionales.

Figura 10

Capacidades de asistentes virtuales mejoradas por LLM



Nota. Obtenido de *Capability Assessment: LLM-Enhanced Chatbots and Rule-Based Systems* (p. 2), por Meduri, 2024, *Revolutionizing Customer Service: The Impact of Large Language Models on Chatbot Performance*

Sin embargo, la implementación de estas tecnologías no está exenta de desafíos. Entre los principales se encuentran la gestión de la privacidad de datos, la precisión de las respuestas en dominios especializados y la escalabilidad en contextos de alta demanda. Según Meduri (2024), el futuro de la atención con IA se orienta hacia capacidades multimodales, reconocimiento de emociones y un servicio predictivo que anticipe necesidades del cliente mediante modelos de comportamiento. Los asistentes virtuales que apliquen IA con LLM multimodales tendrán la capacidad gestionar consultas de clientes basadas en imágenes, proporcionando una experiencia de interacción más completa e intuitiva (p. 8).

3.5 Gestión de Datos Personales en Banca y Gobernanza de IA

3.5.1 Ética Aplicada a la IA

El desarrollo y despliegue de sistemas de inteligencia artificial requiere marcos éticos que orienten tanto su diseño como su uso socialmente responsable. Jobin et al. (2019) identifican que, a pesar de la diversidad de guías elaboradas a nivel internacional, existe convergencia en torno a cinco principios recurrentes: transparencia, justicia y equidad, no maleficencia, responsabilidad y privacidad (p. 7). Estos principios constituyen la base para evaluar la legitimidad de los sistemas de IA y su alineación con valores sociales. En el mismo

sentido, Floridi y Cowls (2019) también añaden que, ante la adopción acelerada de la IA por parte de la sociedad, es necesario analizar y definir los conjuntos de principios éticos más destacados (p. 2), para lo que los autores proponen un marco unificado que adapta cuatro principios bioéticos: beneficencia, no maleficencia, autonomía, justicia y explicabilidad, entendida como combinación de inteligibilidad (capacidad de comprender cómo funciona el sistema) y rendición de cuentas (identificación clara de responsables) (pp. 6-8).

Por su parte, Dignum (2019) complementa esta visión al presentar el enfoque ART de la IA (rendición de cuentas, responsabilidad y transparencia), que establece que la autonomía y adaptabilidad de los sistemas deben estar acompañadas de mecanismos de rendición de cuentas, responsabilidad humana irrenunciable y transparencia verificable (pp. 54-62).

3.5.2 Tratamiento de Datos Personales en Bancos de la Argentina

El tratamiento de datos personales en el sector bancario argentino se encuentra regulado principalmente por la Ley N° 25.326 de Protección de Datos Personales de la República Argentina (2000). La norma define datos personales como “información de cualquier tipo referida a personas físicas o de existencia ideal determinadas o determinables” (Artículo 2), y establece una categoría especial de datos sensibles, entre los que se incluyen origen racial o étnico, opiniones políticas, convicciones religiosas, estado de salud o vida sexual (Artículo 2). El principio de calidad de los datos exige que la información sea veraz, adecuada, pertinente y no excesiva en relación con la finalidad para la que se recolecta (Artículo 4). Asimismo, la ley requiere que los datos se utilicen para fines compatibles con los que motivaron su obtención y que se mantengan actualizados. El consentimiento del titular debe ser previo, expreso e informado, salvo en casos específicos, entre los cuales se encuentran “las operaciones de entidades financieras y la información que reciben de sus clientes” (Artículo 5). La ley también establece el deber de información al momento de la recolección (Artículo 6), la obligación de implementar medidas de seguridad técnicas y organizativas (Artículo 9), y el deber de confidencialidad para quienes accedan a los datos (Artículo 10). Los titulares disponen de los derechos ARCO, compuesto por acceso, rectificación, actualización y supresión (Artículos 14-16). El artículo 20 prohíbe la adopción de decisiones que produzcan efectos jurídicos basadas únicamente en tratamientos automatizados destinados a definir perfiles de personalidad, lo cual resulta directamente relevante frente al uso de algoritmos en banca (Artículo 20).

3.5.3 Marcos Regulatorios y Estándares Aplicables a IA en Banca

La regulación de la IA ha evolucionado hacia un enfoque basado en riesgos, con el objetivo de garantizar que los sistemas se desplieguen de manera segura y confiable. En este marco, el Reglamento (UE) 2024/1689 sobre inteligencia artificial (2024), es una reglamentación conocida como AI Act que establece obligaciones diferenciadas según el nivel de riesgo. Las aplicaciones clasificadas como alto riesgo, entre las que se incluyen los sistemas de evaluación crediticia y de solvencia, deben cumplir con requisitos estrictos de gestión de riesgos, documentación técnica, calidad de los datos, supervisión humana y medidas de ciberseguridad (Artículos 16, 26). El Reglamento también prohíbe prácticas de manipulación subliminal, la puntuación social y el uso de biometría en tiempo real en espacios públicos, salvo excepciones muy limitadas (Artículos 5, 6). Asimismo, exige que los proveedores de sistemas de IA de alto riesgo implementen medidas para prevenir ataques de envenenamiento de datos, ejemplos adversarios y vulnerabilidades en la confidencialidad del modelo (Artículo 76).

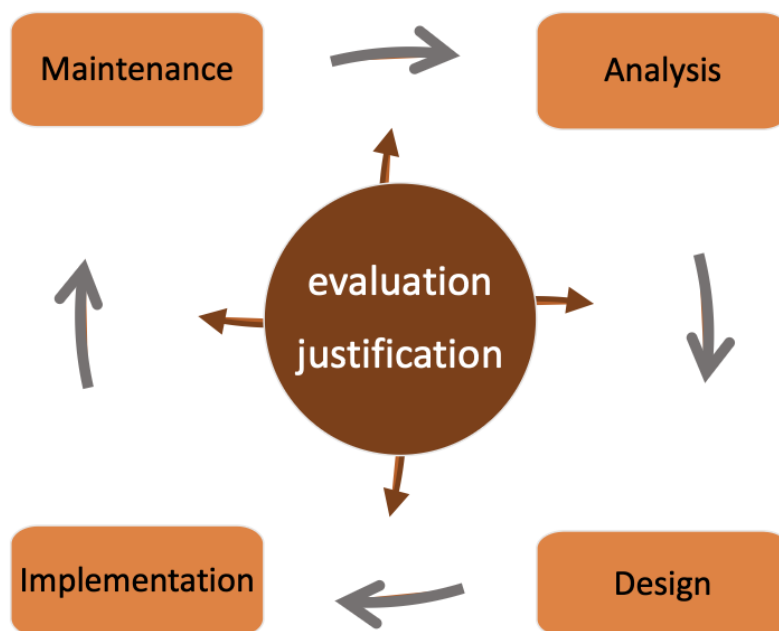
Si bien el AI Act es una norma vigente en el territorio europeo, constituye un estándar de referencia internacional y un punto de comparación para otras jurisdicciones. En el caso de Argentina, puede ser tomado como referencia y guía voluntaria para reforzar las prácticas de gobernanza en el sector bancario, en complemento con la Ley 25.326 y otras normativas locales emitidas por el Banco Central de la República Argentina.

3.5.4 Seguridad, Sesgos y Mitigación de Riesgos en LLM

Los modelos de lenguaje a gran escala presentan riesgos particulares que deben ser mitigados en contextos bancarios. Entre los riesgos técnicos se encuentran los ataques adversarios, el envenenamiento de datos y modelos, y la generación de salidas inexactas o sesgadas. Según Cath (2018), la falta de transparencia y los sesgos algorítmicos pueden comprometer principios de equidad y justicia, proponiendo la auditoría ética como un mecanismo clave para la mitigación de riesgos (p. 2). A su vez, Dignum (2019) señala que el principio de responsabilidad humana indelegable debe garantizarse en el ciclo de vida de los modelos, de modo que las decisiones críticas no recaigan exclusivamente en procesos automatizados (p. 67). Dignum (2019) en la siguiente imagen presenta el modo de lograr un desarrollo responsable desde el punto de vista del ciclo de vida de los sistemas con inteligencia artificial.

Figura 11

Desarrollo responsable en el ciclo de vida de sistemas con IA



Nota. Obtenido de *The Responsible Development Life Cycle for AI systems* (p. 66), por Dignum, 2019, Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way

En Argentina, la Ley 25.326 de Protección de Datos Personales (2000) prohíbe explícitamente la adopción de decisiones basadas únicamente en tratamientos automatizados destinados a evaluar el perfil de las personas, lo cual funciona como un límite normativo directo al uso de LLM en procesos de calificaciones o concesión de crédito sin intervención humana (Artículo 20).

Capítulo 4 – Propuesta de Intervención

4.1 Descripción

La propuesta de intervención se centra en la implementación de un asistente virtual inteligente para el servicio al cliente en banca, diseñado para superar las limitaciones de los chatbots tradicionales. Estos sistemas de primera generación, basados en reglas rígidas, demostraron bajos niveles de satisfacción y generaron frustración al no poder interpretar el contexto de las consultas. De acuerdo con Galileo Financial Technologies y PYMNTS (2025), solo el 29% de los usuarios se declaró conforme con estas herramientas, lo que evidenció su escasa capacidad para resolver necesidades reales de los clientes (pp. 7-9). En contraste, la inteligencia artificial conversacional embebida se integra en los canales digitales con la capacidad de comprender intenciones, procesar lenguaje natural y entregar respuestas personalizadas, transformándose en un recurso estratégico para incrementar la satisfacción y la retención de clientes (Galileo Financial Technologies & PYMNTS, 2025, pp. 9-12).

La necesidad de un modelo más avanzado también se explica por los cambios en el comportamiento de los consumidores financieros. Según Accenture (2025), el 72% de los clientes afirma que la personalización influye en la elección de su banco principal, aunque apenas un 3% utiliza las herramientas que hoy se ofrecen con ese fin (p. 17). Esta brecha genera que muchos usuarios diversifiquen sus relaciones financieras entre distintas instituciones, debilitando la fidelización histórica y obligando a los bancos a diseñar experiencias digitales más cercanas y continuas (Accenture, 2025, pp. 3-8).

El alcance de la propuesta incluye la cobertura de los principales flujos de negocio de la banca minorista. En primer lugar, se encuentran las consultas de información básica como condiciones de productos o consultas relacionadas a sucursales, horarios y canales de atención y también se incluyen operaciones simples como compartir los pasos para realizar transferencias, pago de servicios y recuperación de credenciales, que mejoran la autonomía del cliente y reducen la carga operativa. Un segundo bloque contempla la gestión de reclamos y el seguimiento de casos abiertos para aquellas conversaciones en las que el asistente virtual no pueda resolver de manera autónoma una consulta del cliente con total satisfacción. Como tercer punto, el asistente cumplirá una función preventiva en materia de seguridad digital ofreciendo recomendaciones de protección. El cuarto flujo de resolución propone que el asistente virtual pueda conectarse directamente con repositorios de transacciones y movimientos de cuenta del

banco para resolver incidencias directamente ante situaciones de duplicidad de cargos o explicación a clientes sobre los consumos impactados por transacciones pasadas.

Un aspecto central de la iniciativa es la construcción de confianza con la que debe funcionar el asistente virtual. De acuerdo con World Economic Forum y Accenture (2024), la adopción de inteligencia artificial en servicios financieros depende de la transparencia y la posibilidad de acceder a un asesor humano cuando sea necesario (pp. 18-21). En consecuencia, el diseño del asistente debe garantizar mecanismos claros de derivación hacia operadores humanos en casos sensibles, evitando que la automatización se perciba como una barrera. Esta visión coincide con la tendencia hacia la hiper personalización, identificada como uno de los ejes que transformará el futuro del asesoramiento financiero, ampliando el acceso a servicios de calidad que antes estaban reservados a segmentos de alto patrimonio (World Economic Forum & Accenture, 2024, pp. 22-27).

4.2 Metodología

Este trabajo adopta una metodología mixta, ya que se adapta como la mejor forma de presentar la propuesta de intervención aplicada al ámbito tecnológico. Según Hernández Sampieri (2014), “la meta de la investigación mixta no es reemplazar a la investigación cuantitativa ni a la investigación cualitativa, sino utilizar las fortalezas de ambos tipos de indagación, combinándolas y tratando de minimizar sus debilidades potenciales” (p. 532).

Hernández Sampieri (2014) expresa que la investigación mixta “se fundamenta en el pragmatismo”, utilizando la recolección de datos, el análisis de los datos hallados y una posterior integración de inferencias cuantitativas y cualitativas (p. 533). El enfoque cuantitativo será utilizado para obtener conclusiones sólidas con respecto a las necesidades de los centros de atención al cliente y el funcionamiento de los modelos de lenguaje a gran escala, consolidando un marco teórico que respalde la propuesta de implementación. Con respecto a los datos cualitativos, como por ejemplo el análisis de los resultados del desarrollo del prototipo, brindará un apoyo práctico a las soluciones propuestas. Según Hernández Sampieri (2014) el principal aporte de la metodología mixta es que permite la utilización de datos cuantitativos y cualitativos en conjunto, es decir, que son multi metódicos (p. 535).

De acuerdo con Hernández Sampieri (2014), el enfoque mixto permite “obtener una visión más comprensiva sobre el planteamiento si se emplean ambos métodos, así como un mayor entendimiento, ilustración o clarificación de los resultados de un método sobre la base de los resultados del otro” (p. 538), y brinda la posibilidad “proveer un contexto más completo,

profundo y amplio, pero al mismo tiempo generalizable y con validez externa” (Hernández Sampieri, 2014, p. 539).

4.3 Plan de implementación del asistente virtual inteligente

4.3.1 Alcance Funcional

Agentforce, el constructor de Salesforce utilizado para el desarrollo del asistente virtual, permite construir agentes capaces de razonar, decidir y actuar sobre datos empresariales dentro de un entorno seguro e integrado. Según Kihn (2025), esta tecnología combina la potencia de los modelos de lenguaje a gran escala con la estructura de datos y la lógica de negocio del ecosistema Salesforce, lo que posibilita automatizar procesos de negocio a partir de la toma de decisiones (p. 21).

De acuerdo con la Agentforce Guide (Salesforce, 2025a), la definición del alcance debe comenzar por identificar las tareas específicas que el agente resolverá, conocidas como “Jobs to Be Done”, es decir, las tareas o procesos que serán realizados. Al iniciar el diseño del agente, se deben definir los temas base que el asistente virtual podrá interpretar y luego las acciones que permiten ejecutar automatizaciones de acuerdo con las instrucciones que fueron previamente definidas para el asistente virtual. Por ejemplo, si un cliente envía un mensaje, el asistente virtual primero determinará a qué tema se relaciona el mensaje ingresante y luego buscará las acciones que lo asistirán para resolver lo que el cliente está solicitando (párr. 16).

En esta propuesta, el asistente virtual se enfocará en cuatro flujos digitales de un banco: en primer lugar buscará resolver las consultas generales sobre cuentas financieras y dudas relacionadas a cómo operar con el banco; en segundo lugar, se incluye la gestión de reclamos, con la capacidad de recopilar datos del cliente, generar un caso en el módulo de gestión de reclamos Service Cloud y escalar la solicitud a un agente humano cuando la situación lo requiera; en tercer lugar se propone incluir un módulo de seguridad y accesibilidad en el que los clientes puedan consultar sobre los pasos que deben seguir ante posibles fraudes, recomendaciones y la posibilidad de gestionar una denuncia. En cuarto lugar, se consideran las operaciones de autogestión sobre movimientos de la cuenta, escenario de mayor complejidad de implementación porque se involucrarán a los sistemas internos del banco que son responsables de validar la identidad de las personas y de compartir información en línea sobre saldos, movimientos y transacciones pasadas.

El proyecto contempla una primera fase con un alcance acotado para permitir a los bancos una rápida implementación que remplace los chatbots actuales con una mejora

significativa en la experiencia de usuario. El asistente no ejecutará operaciones monetarias directas, transferencias entre cuentas ni procesos que involucren acciones que no sean sólo de consulta sobre los activos financieros de los clientes. Gotts et al. (2024) sostienen que los proyectos más exitosos con Agentforce comienzan con flujos transaccionales simples, con la capacidad de trasladar la comunicación a un humano siempre que sea necesario. Para evitar que toda conversación derive en una asistencia de una persona, los agentes deben ser eficaces, es decir, que deben ser bien instruidos, comprender los procesos que el cliente requiere y trabajar con datos limpios (p. 9).

4.3.1.1 Flujos de procesos de atención al cliente

Para lograr la operatividad del alcance funcional se diseñaron cuatro flujos de procesos que representan los recorridos iniciales de interacción entre el cliente y el asistente virtual. Cada flujo refleja una situación típica de atención en banca minorista y detalla las etapas que atraviesa la consulta desde su inicio en el canal digital hasta la resolución final o la derivación a un agente humano. Estos flujos permiten visualizar la lógica operativa del sistema y constituyen una guía recomendada que podrá ser modificada o ampliada por cada banco en particular en caso de ser necesario.

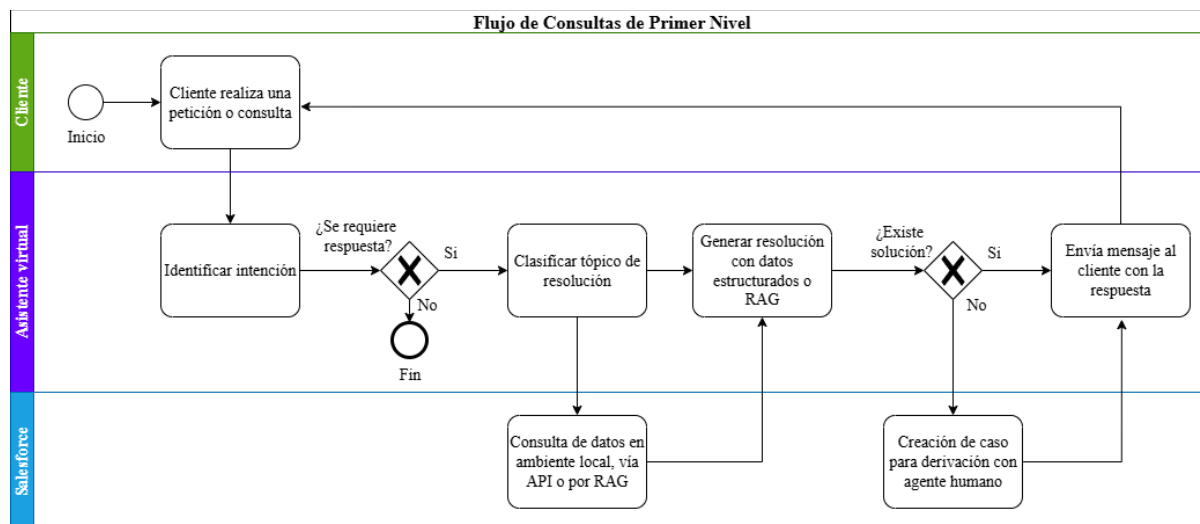
Flujo de Consultas de Primer Nivel. Este proceso abarca las interacciones iniciales de los clientes que buscan información general, como la verificación de saldos, la consulta de condiciones de productos financieros o la localización de sucursales y horarios de atención. El proceso se inicia cuando el usuario ingresa su mensaje en el canal digital en el sitio web del banco. El asistente identifica la intención mediante procesamiento de lenguaje natural y la asocia al tópico “Consultas básicas”. A partir de esta clasificación, el agente ejecuta una acción de búsqueda en Data Cloud que funciona como un repositorio documental del banco, recuperando la información relevante. La respuesta se entrega al cliente en lenguaje natural y con formato breve, pudiendo incluir vínculos o instrucciones complementarias. Si el usuario requiere más detalles, el flujo continúa con una nueva consulta dentro del mismo tópico; de lo contrario, el sistema registra el cierre automático de la interacción.

También se incluyen en este flujo aquellas operaciones de autogestión que no implican movimientos monetarios directos, como las instrucciones para realizar transferencias, pagos de servicios o recuperación de credenciales. El cliente inicia la interacción solicitando asistencia para ejecutar una operación. El asistente identifica la intención y selecciona la secuencia de pasos correspondiente, guiando al usuario con mensajes ordenados. El flujo concluye con la

confirmación al usuario de que el procedimiento fue completado o, en caso de error, con la generación de un caso en Service Cloud para revisión por parte de un asistente humano.

Figura 12

Diagrama BPMN para Consultas Básicas

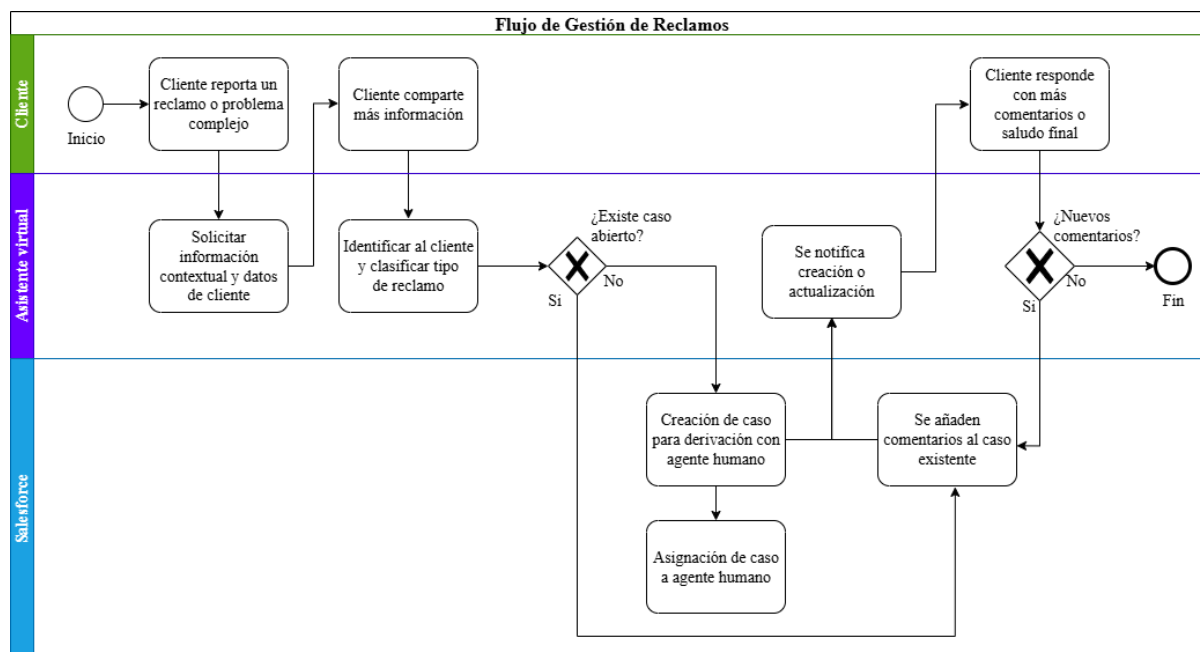


Nota. Elaboración propia con la herramienta draw.io y utilizando las formas provistas por el Modelo y Notación de Procesos de Negocio (BPMN) 2.0.

Flujo de Gestión de Reclamos. Este flujo contempla situaciones en las que el cliente reporta un problema que debe registrarse en el sistema para su correspondiente gestión o solicita seguimiento de un caso abierto previamente. El proceso se inicia cuando el usuario comunica un inconveniente o expresa insatisfacción. El asistente virtual analiza el mensaje, clasifica el tipo de reclamo y recopila los datos básicos necesarios como cliente, asunto del problema, descripción del caso y fecha y hora del incidente. Con esta información, el agente crea automáticamente un registro en Service Cloud mediante una acción predefinida, asignándole un número de seguimiento y categoría prioritaria. El cliente recibe una confirmación inmediata. En reclamos de baja complejidad, el sistema puede ofrecer soluciones directas, mientras que los casos de mayor sensibilidad se derivan a un operador humano del área de atención al cliente.

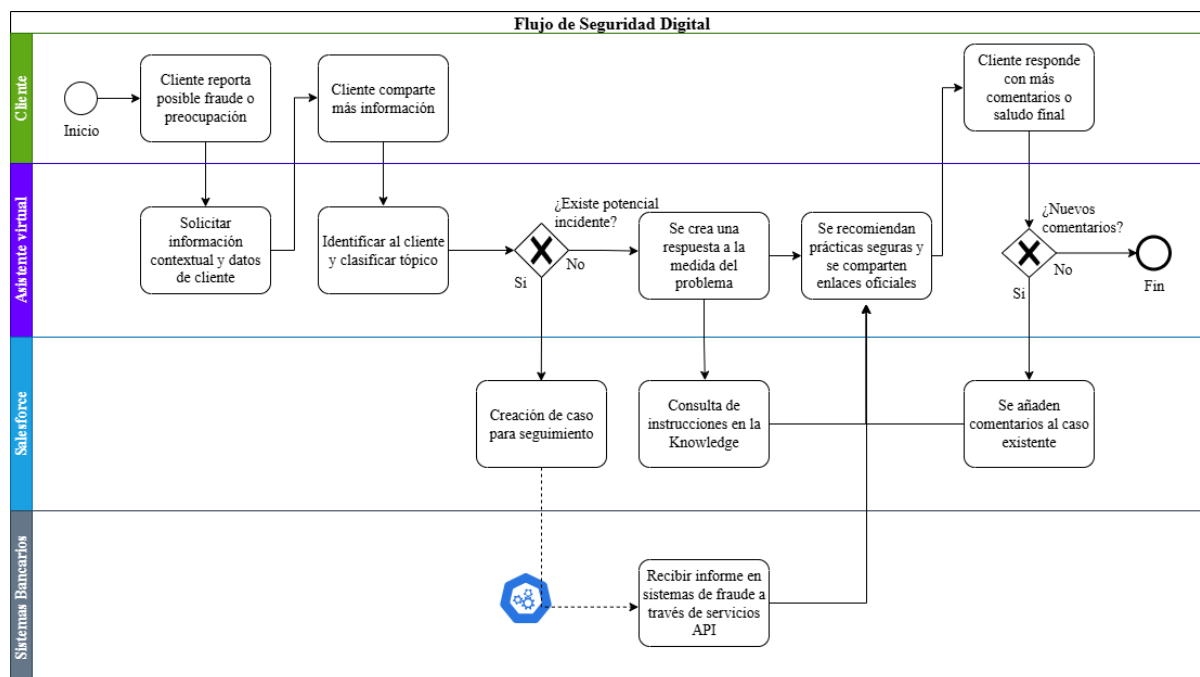
Figura 13

Diagrama BPMN para Gestión de Reclamos



Nota. Elaboración propia con la herramienta draw.io y utilizando las formas provistas por el Modelo y Notación de Procesos de Negocio (BPMN) 2.0.

Flujo de Seguridad Digital. El tercer flujo responde a la necesidad de reforzar la educación y la prevención en materia de ciberseguridad, un aspecto crítico en el sector financiero. El asistente virtual actúa de manera reactiva y preventiva. En modo reactivo, responde a consultas relacionadas con contraseñas, fraudes o phishing, brindando instrucciones seguras y actualizadas según las políticas del banco. En modo preventivo, puede emitir alertas automáticas cuando detecta patrones de conversación vinculados con riesgos potenciales o intentos de estafa que pueden haber sufrido los clientes, brindando la posibilidad de conectarse de manera urgente con un asistente humano y ofreciendo los canales de atención más urgentes como la línea telefónica del banco o de empresas de los productos asociados como tarjetas de crédito.

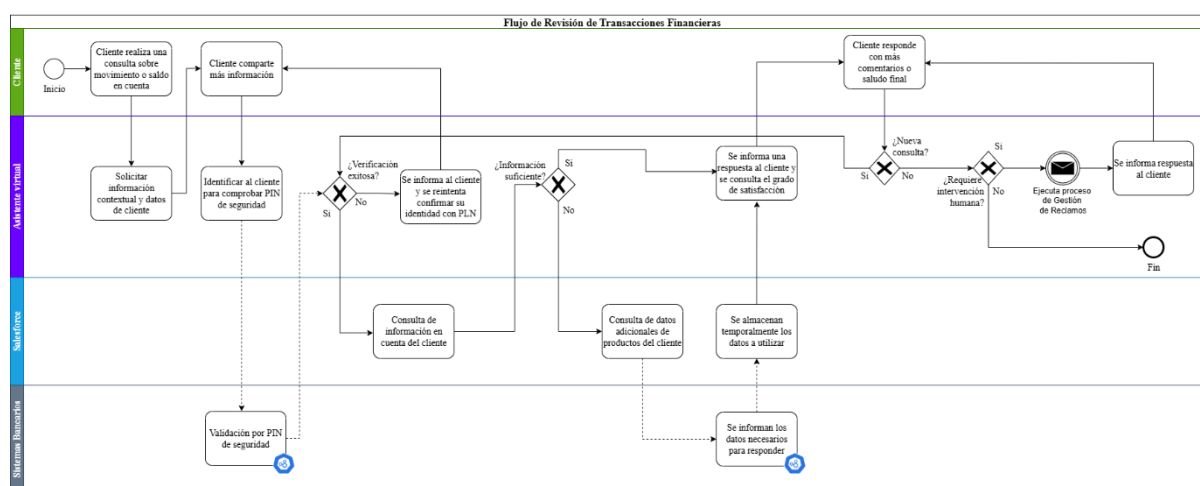
Figura 14**Diagrama BPMN para Seguridad Digital**

Nota. Elaboración propia con la herramienta draw.io y utilizando las formas provistas por el Modelo y Notación de Procesos de Negocio (BPMN) 2.0.

Flujo de Revisión de Transacciones Financieras. El flujo aborda interacciones de mayor complejidad, en las que el cliente requiere revisar movimientos, detectar duplicidades de pago o comprender cargos en su cuenta. Representa un caso avanzado de integración entre el asistente virtual y los sistemas transaccionales del banco, donde la exactitud y la trazabilidad de la información resultan esenciales. El proceso se inicia cuando el usuario consulta sobre un movimiento específico o manifiesta una discrepancia en su saldo. El asistente analiza la intención mediante procesamiento de lenguaje natural y determina si se trata de una revisión de transacciones, un reclamo por duplicidad o un cargo no reconocido. A continuación, ejecuta una consulta segura en los sistemas internos del banco o en registros accesibles por Data Cloud, empleando autenticación del cliente y mecanismos RAG para contextualizar la respuesta con datos y políticas internas.

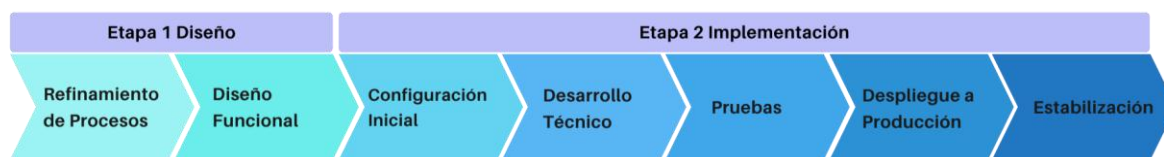
Si la información se obtiene correctamente como para dar una respuesta sólida, el asistente presenta el detalle de la transacción (por ejemplo: fecha, importe, canal de origen) y ofrece una explicación basada en las reglas del banco. Cuando el análisis evidencia una duplicidad o error, se genera automáticamente un caso en Service Cloud, clasificando el incidente como “Revisión de movimiento” y asignando prioridad media o alta según impacto

Diagrama BPMN para Revisión de Transacciones Financieras



4.3.2 Etapas de Implementación

La implementación del asistente virtual con los flujos mencionados y la aplicación de la tecnología Salesforce se estructurará en dos etapas, una orientada al diseño de la solución en conjunto con el banco y una segunda etapa de implementación en la que se configurará y desarrollará la solución en Salesforce, incluyendo la puesta en producción del sistema y la capacitación a usuarios claves. Dentro de las etapas se despliegan unas siete actividades generales sucesivas que permiten avanzar de manera controlada, garantizando la calidad del modelo, la seguridad de los datos y la alineación con los objetivos del banco. En la siguiente imagen se presentan las etapas de ejecución necesarias para la implementación del asistente virtual.

Figura 16**Etapas de Implementación del Asistente Virtual**

Nota. Elaboración propia con la herramienta Drawio.

La implementación comenzará en la segunda fase del proyecto con la configuración propia del asistente virtual (agente) bajo un modelo práctico de implementación adaptado de acuerdo con lo propuesto por Gotts et al. (2024). De esta manera se proponen los siguientes pasos: Refinamiento de Procesos y Sistemas, Diseño Funcional, Configuración Inicial, Configuración Técnica, Pruebas y Despliegue a Producción (pp. 21-29). A continuación, se detalla cada una de las etapas:

1. Refinamiento de Procesos

La primera etapa del proyecto contendrá la ejecución de un refinamiento de procesos con los actores más importantes del banco, para comprender la situación actual del banco con respecto a los procesos de negocio y el ecosistema de aplicaciones con las que se cuenta en la organización. Esta etapa no busca un rediseño completo de procesos, sino que se realizará un conocimiento previo de los procesos del banco y luego un refinamiento de los casos de uso que se proponen en el asistente virtual para que sea posible adaptar su funcionamiento a lo requerido por cada negocio bancario en particular. Para realizar el relevamiento con los stakeholders, se propone una serie de reuniones con los actores principales del banco, lo que permitirá refinar los procesos y finalmente obtener casos de uso para que ejecute el asistente virtual en el centro de atención al cliente.

Según García Molina et al. (2006), la finalidad de relevar y modelar el funcionamiento de un negocio es describir un proceso, especificando sus datos, tareas, roles y reglas de negocio:

Para conseguir sus objetivos, una empresa organiza su actividad por medio de un conjunto de procesos de negocio. Cada uno de ellos se caracteriza por una colección de datos que son producidos y manipulados mediante un conjunto de tareas, en las que ciertos agentes (por ejemplo, trabajadores o departamentos) participan de acuerdo a un flujo de trabajo determinado. Además, estos procesos se hallan sujetos a un conjunto de reglas de negocio, que determinan las políticas y la estructura de la información de la empresa (p. 4).

Una vez definidos los procesos, será posible comenzar a definir los casos de uso del negocio con los roles (actores) involucrados para cada tarea dentro de un mismo proceso (García Molina et al., 2006, p. 5).

2. Diseño Funcional

Con la documentación de los procesos de negocio que se realizó durante la primera actividad general, el equipo de trabajo determina qué procesos serán automatizados y cómo se alinean con los objetivos de negocio, siendo ideal comenzar por aquellos casos de uso que son más simples y fáciles de implementar (Gotts et al., 2024, p. 23). Esta etapa será fundamental para definir los diseños de componentes a implementar y asegurar que los casos de uso acordados con el banco coinciden con los planteados en la propuesta, incluso en tiempos de implementación y fases necesarias.

3. Configuración Inicial

Como lo remarcen Gotts et al. (2024), en esta etapa se prepara el entorno técnico de Salesforce necesario para construir el agente y la activación de la IA Generativa. Se configuran los permisos de usuario, las credenciales de acceso, las integraciones con Data Cloud y los componentes de seguridad del Einstein Trust Layer (p. 22). Además, se definen los lineamientos generales de gobernanza y auditoría que permitirán monitorear el desempeño del agente. Según Salesforce (2025b), una planificación sólida en esta instancia es clave para garantizar que el proyecto cumpla con los requisitos esperados en la definición de casos de uso y el mapa de procesos (**párr. 5**).

4. Desarrollo Técnico

Gotts et al. (2024) indican que en esta etapa se definen los tópicos, las instrucciones y las acciones que conformarán el comportamiento del agente dentro del *Atlas Reasoning Engine* (motor de razonamiento Atlas). Cada tópico representa un grupo de consultas específicas, las instrucciones determinan cómo debe responder el agente, y las acciones definen las tareas que puede ejecutar, como crear un caso con una consulta que debe supervisar una persona del banco o recuperar información desde el repositorio centralizado de clientes en Data Cloud. En esta etapa se crea el agente utilizando Agentforce Builder, y se redactan los mensajes de bienvenida, error y cierre, siguiendo el estilo comunicacional del banco, para que el asistente virtual pueda desarrollarse con un lenguaje aproximado lo más posible al humano (pp. 24-25). Luego, se desarrollarán los flujos incluidos como parte de los procesos de negocio que debe resolver el asistente virtual de manera autónoma. Se configuran los flujos de automatización, las integraciones con sistemas internos y las variables de contexto que permitirán personalizar las

respuestas (Gotts et al., 2024, p. 26). Según Kihn (2025), esta configuración será la que le permita al asistente virtual reclasificar cada interacción con un cliente en un proceso de transformación de los mensajes ingresantes en un tema o tópico con el que puede trabajar, para pasar así a definir cuáles son las acciones que tiene disponible el agente en esa interacción (p. 39).

5. Pruebas

Según Salesforce (2025b), una vez construido el agente puede ser sometido a pruebas manuales directamente en *Agent Builder* (constructor de agente) para verificar su precisión, estabilidad y coherencia. Otra opción es realizar pruebas por lotes utilizando el centro de pruebas denominado *Agentforce Testing Center* (centro de pruebas de Agentforce). Los resultados de las pruebas podrán ser documentados y se utilizarán para ajustar los parámetros del modelo antes de su puesta en producción (párr. 34). Es una de las etapas más trascendentales de la implementación porque si las pruebas no superan un alto porcentaje de aceptación, puedo ponerse en riesgo la posibilidad de desplegar al asistente virtual de cara a los clientes.

6. Despliegue a Producción

Una vez validadas las funcionalidades, el agente se implementa en los canales digitales del banco, inicialmente en una modalidad de Producto Mínimo Viable para un grupo de usuarios o clientes que puedan utilizar las capacidades del asistente virtual y brinden una devolución a partir de su experiencia. Como lo detalle Gotts et al. (2024), en esta etapa se incluye la migración de los metadatos de un ambiente de desarrollo y pruebas a un entorno productivo. A continuación, se realiza una migración de datos y finalmente se comunican los a los equipos de trabajo para comenzar a trabajar con el sistema (p. 27). Al final de la etapa serán capacitados los equipos internos del banco que utilizarán Salesforce como herramienta de gestión de consulta de los clientes y se comunicará a los clientes seleccionados para las pruebas iniciales del asistente virtual.

7. Estabilización

La etapa de estabilización corresponde al cierre del proyecto en el que los equipos técnicos del banco quedan a disposición de monitorear las primeras respuestas del agente en tiempo real, su consumo y las posibles incidencias técnicas que podrían suceder, a efectos de garantizar la operatividad del sistema con normalidad.

4.3.3 Arquitectura Técnica y Componentes

La arquitectura técnica del asistente virtual propuesto se apoya en la infraestructura nativa de Salesforce Agentforce, la cual combina modelos de lenguaje de gran escala con los

datos y procesos empresariales del ecosistema Salesforce. Este enfoque permite que el agente no solo genere respuestas, sino que también pueda simular el razonamiento humano, ejecutar acciones y mantener un control seguro sobre la información del cliente. Según Kihn (2025), los agentes desarrollados con esta tecnología operan como una extensión de la fuerza laboral humana, simplificando la forma en la que las empresas ejecutan sus procesos (p. 22).

En su núcleo, el agente se basará en el Atlas Reasoning Engine, un motor de razonamiento diseñado para interpretar la intención del usuario, seleccionar el tema adecuado y decidir qué acción realizar a continuación. La Agentforce Guide (Salesforce, 2025a) explica que este proceso ocurre a través de tres elementos fundamentales: tópicos, instrucciones y acciones (párr. 12).

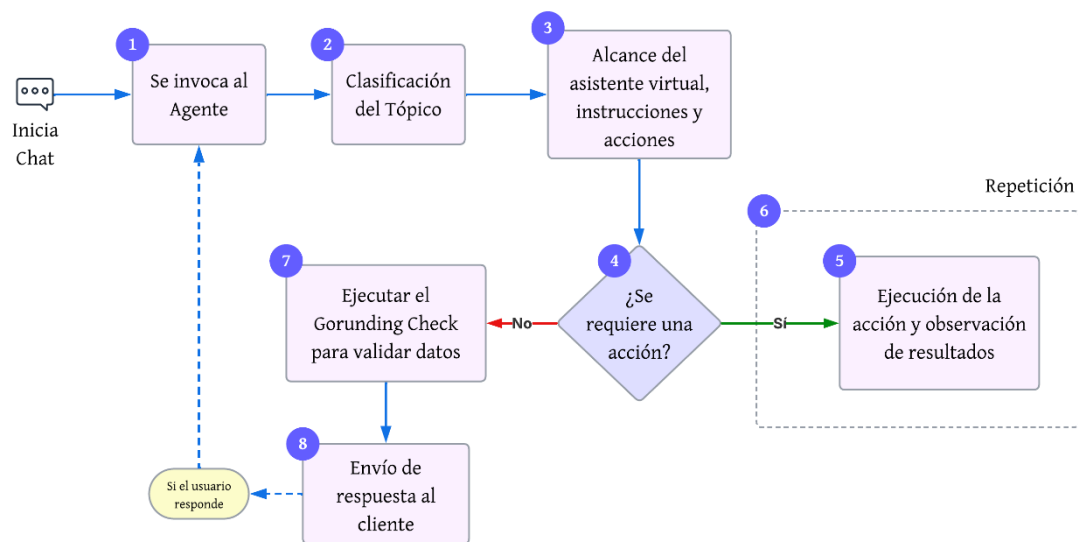
De acuerdo con la Guía de Agentforce, Salesforce (2025a) define a los tópicos como las competencias o dominios del agente, es decir, qué tipo de consultas puede atender. Cada tópico tiene un alcance específico que delimita sus funciones y evita solapamientos. Las instrucciones por su parte establecen las pautas de conversación y los criterios que guían las decisiones del agente dentro de cada tópico, permitiendo mantener coherencia en el tono, la secuencia de pasos y la forma en que el sistema solicita o presenta información. Y las acciones representan las tareas que el agente puede ejecutar, como buscar datos en Data Cloud, cambiar el valor de un atributo de un cliente, o registrar un caso para ser derivado a una persona en el módulo de atención al cliente Service Cloud. Las acciones pueden basarse en reglas de bajo código a partir de la creación de flujos en la herramienta Flow, o código avanzado en el lenguaje Apex, lo cual requiere una mayor dedicación, pero permite aplicar más complejidad en la operación. Estos tres componentes trabajan en conjunto para permitir que el agente razone de manera estructurada. El motor analiza el mensaje del usuario, identifica el tópico correspondiente, aplica las instrucciones del contexto y, si es necesario, ejecuta una acción. Una vez completada la tarea, la respuesta se valida mediante un proceso de verificación conocido como *grounding check* (verificación de anclaje), que asegura que la información esté respaldada por datos reales del sistema. Esta lógica reduce el riesgo de respuestas imprecisas o alucinaciones del modelo. Un punto destacable de este funcionamiento es que si el usuario saluda con un mensaje como “Hola” o cualquier otro mensaje que no requiera ejecutar un proceso de razonamiento complejo, el asistente virtual no deberá ejecutar ninguna acción porque se encuentra preparado para brindar una respuesta rápida de acuerdo con el contexto, lo que brinda una novedosa capacidad de adaptarse al contexto. Luego, si ese mismo cliente que está interactuando con el asistente virtual desea por ejemplo consultar los horarios de

atención en una sucursal del banco, entonces el asistente virtual en este caso tendrá que ejecutar su primera acción para ir a buscar en su base de datos esa información, validarla y finalmente entregársela al cliente como respuesta (Salesforce, 2025a).

En la siguiente imagen se presenta en detalle el funcionamiento del motor Atlas Reasoning Engine, el cual permite ejecutar el proceso detallado anteriormente.

Figura 17

Motor de Razonamiento Atlas



Nota. Elaboración propia con la herramienta Lucidchart. Adaptado de *The Atlas Reasoning Engine*, por Salesforce, 2025a, Introducing the Agentforce Guide.

Como se visualiza en la imagen, el proceso inicia con un requerimiento que un cliente realiza al asistente virtual por chat (indicado con el número 1 en la imagen). Luego de recibir el mensaje del cliente, este se clasifica para determinar el tema (tópico) al que corresponde (2), lo que determinará que instrucciones y acciones podrá ejecutar el agente en ese caso (3). El motor de razonamiento elige si debe ejecutar una acción o no de acuerdo con lo preguntado por el cliente (4). Si debe hacerlo, entonces ejecuta la acción tantas veces como sea necesario o incluso con nuevas acciones posteriores (5 y 6). Se evalúa la respuesta generada por el agente y luego se asegura que cumpla con los estándares de comunicación (7). Finalmente se envía el mensaje al cliente y el asistente virtual queda a la espera para volver a comenzar el mismo proceso si el cliente realiza una respuesta al mensaje anterior (8).

Según Kihn (2025), el Atlas Reasoning Engine actúa como el cerebro de Agentforce, combinando modelos, datos y flujos empresariales en una arquitectura unificada (p. 89). El autor complementa que el motor Atlas fue concebido con una arquitectura modular compuesta por servicios técnicos independientes, cada uno especializado en funciones específicas como el procesamiento del lenguaje, la recuperación de información o la generación de respuestas. Esta estructura le permite operar de forma flexible y eficiente frente a grandes volúmenes de datos y en entornos distribuidos globalmente. Dado que Salesforce atiende a corporaciones multinacionales con altos requerimientos de seguridad, el diseño de Atlas incorporó mecanismos de autenticación, autorización y protección de la información. Para garantizar su confiabilidad, Atlas fue creado de manera controlada y progresiva dentro de la infraestructura Hyperforce, siendo probado inicialmente por un grupo reducido de clientes con exigencias críticas de seguridad (p. 98). Según Kihn (2025), la infraestructura Hyperforce surgió en Salesforce como la respuesta directa a la evolución de los grandes proveedores de infraestructura en la nube, conocidos como hyperscalers, que ofrecen mayor flexibilidad y eficiencia en la gestión de recursos. En sus inicios, Salesforce mantenía sus propios centros de datos, lo que implicaba altos costos y limitaciones operativas. Con el crecimiento de los hyperscalers, capaces de ofrecer infraestructuras distribuidas globalmente, la compañía desarrolló Hyperforce como una plataforma adaptable diseñada para operar sobre distintos proveedores de nube en más de veinte regiones. Este enfoque permite cumplir con los requerimientos de residencia de datos impuestos por países como India, Canadá, Japón o Alemania, garantizando que la información de los clientes se almacene físicamente dentro de su jurisdicción (pp. 57-58).

Otro componente esencial de la arquitectura mencionado por Kihn (2025) es Salesforce Data Cloud, que actúa como repositorio de datos y sistema de indexación para la recuperación de información en tiempo real sin programación en caso de que se trabaje con archivos del banco que serán utilizados como base de conocimiento (p. 58), y en caso que los bancos se nieguen a mover sus datos de clientes hacia Salesforce, existe una alternativa conocida como copia-cero, una funcionalidad de Data Cloud que le permite a las empresas incorporar datos de bases de datos externas a la plataforma sin la necesidad de realizar una persistencia de datos, por ejemplo recolectando y utilizando datos en tiempo real de otros sistemas como Snowflake, Databricks, Google o Amazon (p. 59). Con respecto a la interacción que el asistente virtual tendrá con los registros de datos del cliente, operaciones recientes o estados de cuenta, se utilizará el método conocido como Retrieval-Augmented Generation o Generación de

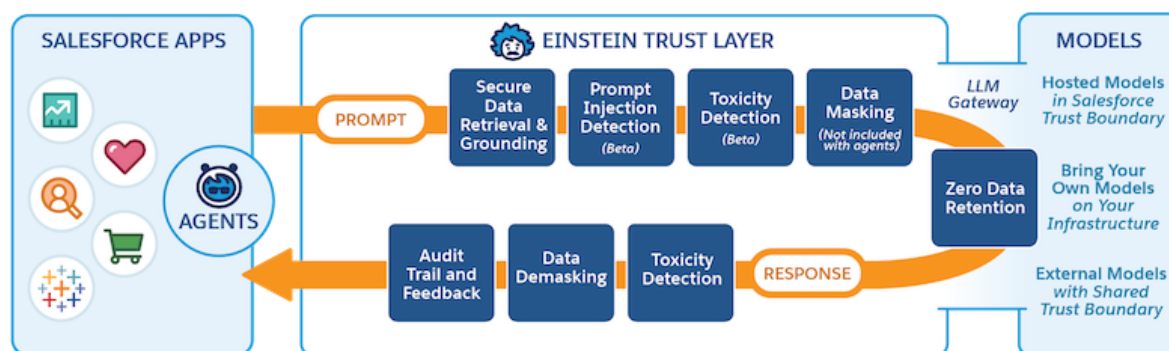
Recuperación Aumentada (RAG, por sus siglas en inglés), que combina la consulta a bases estructuradas con la generación contextual de respuestas a partir de la consulta de datos no estructurados (Kihn, 2025, p. 81).

Para asegurar que el asistente virtual responda de forma confiable antes de su puesta en producción, se propone la utilización de Agentforce Testing Center de Salesforce, una herramienta diseñada para probar agentes de IA. Este centro de pruebas permite generar escenarios con casos de interacción en los que se pruebe la ejecución en lotes, recibir informes detallados de desempeño (por ejemplo, qué porcentaje de los casos fallaron) y ajustar iterativamente el agente con base en esos resultados. A su vez, Testing Center puede complementarse con la Testing API y herramientas de desarrollo como la CLI de Agentforce (Agentforce DX) para automatizar esas pruebas dentro de pipelines de desarrollo, lo que brinda una gran capacidad para todo tipo de flujos (Help Salesforce, 2025a).

Otro componente clave en la implementación será la utilización de la capa conocida como Einstein Trust Layer, la cual funciona como una interfaz intermedia entre los datos del banco alojados en la plataforma Salesforce y los modelos de lenguaje externos, como GPT, Claude o Gemini (y que en la propuesta se propone construir con el modelo GPT). Todas las solicitudes que se ejecutan desde Salesforce hacia estos modelos pasan por el LLM Gateway, que incluye esta capa de confianza. Su propósito es actuar como un filtro inteligente y un árbitro ético, protegiendo la información sensible del cliente y garantizando el cumplimiento de las políticas de privacidad.

Figura 18

Funcionamiento de la Capa de Confianza Einstein



Nota. Obtenido de *Einstein Trust Layer: Designed for Trust*, por Salesforce, s.f.b.

Como se presenta en la imagen, las principales funciones de la Einstein Trust Layer son la recuperación segura de datos, que impide la fuga o exposición de información personal

identificable (PII, por sus siglas en inglés), y el anclaje dinámico conocido como *dynamic grounding* (anclaje dinámico), que permite incorporar datos actualizados del cliente en tiempo real dentro del prompt enviado al modelo. Un caso práctico de esta aplicación sería por ejemplo si se requiere compartir con el LLM el nombre de un cliente que se llama “Carlos”, en lugar de decirle al LLM que nuestro cliente es “Carlos” se le aplica una máscara que permite cambiar ese nombre por ejemplo por una etiqueta como “Cliente1”, para que el modelo no acceda a datos confidenciales. Luego la política de cero retenciones de datos asegura que la información no sea almacenada ni reutilizada por terceros, la detección de toxicidad se encarga de filtrar contenidos inapropiados o sesgados en las entradas y salidas, y finalmente, la trazabilidad y registro de auditoría, mantiene un historial verificable de las interacciones para garantizar transparencia y responsabilidad en el uso de la IA. Estas funcionalidades están incorporadas de manera nativa en el agente y tienen conexión directa con los datos de clientes que se almacenan en las “Salesforce Apps” que se visualiza en la imagen (Kihn, 2025, p. 61).

La arquitectura propuesta, por tanto, combina razonamiento autónomo, integración con datos empresariales y control de seguridad. El agente funciona dentro de un entorno cerrado, con acceso limitado a la información que requiere para resolver consultas definidas. Esta estructura garantiza que las decisiones del modelo estén sustentadas en datos verificables, preservando al mismo tiempo la trazabilidad y la transparencia del proceso.

En la siguiente imagen se presenta la arquitectura final de componentes que tendrá la solución, con la activación de Data Cloud como módulo de control y gestión de datos (estructurados y no estructurados), el módulo de Servicio que será el sistema back-end que funcione como soporte para la resolución de consultas complejas que deberán ser atendidas por una persona, incluida la generación de una base de conocimiento del banco y la creación de consultas para ser atendidas fuera de línea, y la implementación del asistente virtual representado en Agentforce como herramienta principal de inteligencia artificial con conexión hacia el modelo de lenguaje GPT.

Figura 19

Arquitectura de Componentes de la Solución



Nota. Elaboración propia con la herramienta Lucidchart.

Los servicios de plataforma hacen referencia a capacidades de la plataforma de Salesforce que ya brinda un sistema completamente seguro para que los usuarios del banco ingresen de manera remota por ser una tecnología en la nube que sólo requiere de un dispositivo con acceso a Internet. De izquierda a derecha se visualizan las capacidades de contar con una infraestructura de datos en la nube, ser una plataforma de datos de clientes, personalización, capacidad de administrar la plataforma, autenticación de usuarios que utilizan el sistema, rastreo de contactos del cliente y analítica con reportes propios.

4.3.4 Integración con procesos bancarios

La efectividad del asistente virtual depende de su capacidad para integrarse de forma fluida con los procesos y sistemas centrales del banco. Salesforce Agentforce permite que el agente no solo interactúe con los clientes, sino que también ejecute acciones sobre los datos reales del ecosistema institucional, garantizando que cada respuesta esté respaldada por información actualizada y segura.

El agente podrá interactuar con datos de clientes existentes en bases de datos del banco. La capacidad de Salesforce como plataforma web de conectarse con sistemas externos permitirá que se desarrollen integraciones punto a punto entre un sistema externo del banco y la plataforma, para que el asistente virtual pueda consultar datos en tiempo real o informar hacia otro sistema algún dato, como por ejemplo notificar a un sistema central del banco que

un cliente realizó una denuncia de la pérdida de una tarjeta de crédito en una conversación con el agente virtual.

Con respecto a la utilización de datos históricos, se propone la activación de la herramienta Data Loader provista de manera gratuita por la plataforma Salesforce para realizar una migración inicial de clientes y entidades relacionadas al cliente. La correcta migración de datos constituye un requisito esencial para el funcionamiento del asistente virtual, dado que el modelo requiere acceder a información real y actualizada para generar respuestas precisas. Data Loader es una herramienta de uso local que puede descargarse para utilizar en sistemas operativos como Windows o macOS y permite importar hasta 150 millones de registros con datos de manera controlada a partir de la creación de archivo tipo .csv que deberán ser compartido por el banco (Salesforce, s.f.e). Este proceso implica mapear los campos del sistema de origen con los objetos estándar de Salesforce, tales como Account (cuentas de clientes del banco), Financial Account (cuentas financieras relacionadas al cliente) o Case (reclamos relacionados al cliente). En caso de que existan datos específicos del banco que no tengan correspondencia directa con los objetos nativos, se podrán crear objetos personalizados (Custom Objects) dentro de la organización Salesforce, asegurando así la preservación de toda la información relevante. Por ejemplo, este podría ser el caso de una entidad en la que el banco guarde interacciones previas con el cliente que deberán ser utilizadas como histórico de relacionamiento.

Otro nivel de integración se desarrollará sobre Salesforce Service Cloud, que actúa como el sistema central de gestión de casos (consultas o reclamos de clientes). Según la Agentforce Guide (Salesforce, 2025a), cuando un cliente realiza una consulta o presenta un reclamo, el asistente puede crear automáticamente un registro con los datos esenciales del contacto, tipo de solicitud y nivel de prioridad. Esta automatización permite estandarizar el flujo de atención y reducir la intervención manual. Las acciones de creación o actualización de casos se ejecutan a través de acciones diseñadas dentro del agente de atención.

Además, se propone una integración con los canales digitales de atención, incluyendo como posibles canales dentro del alcance de implementación la disponibilidad del asistente virtual en la página web del banco. El agente se implementará inicialmente en el canal web del banco, donde atenderá consultas recibidas en tiempo real. La conexión se realiza mediante una interfaz de mensajería embebida en el sitio, que se comunica con el entorno de Salesforce. En etapas posteriores, podrá ampliarse al canal de mensajería móvil y otros canales como

WhatsApp, garantizando una experiencia de omnicanalidad, dado que la funcionalidad permite sumar nuevos canales de comunicación sin un gran esfuerzo de desarrollo adicional.

4.4 Recursos

4.4.1 Recursos Humanos

La ejecución del proyecto requiere la participación de un equipo interdisciplinario conformado por profesionales técnicos, analistas de negocio y funcionales, y responsables de gestión. Cada rol cumple una función específica dentro del ciclo de vida del asistente virtual, asegurando que la solución se diseñe, implemente y mantenga conforme a los estándares de calidad del banco y las buenas prácticas de Salesforce. De acuerdo con lo recomendado por Project Management Institute (2013), se planificará un equipo dedicado, lo que significa que “la asignación de todos o de la mayor parte de los miembros del equipo del proyecto es para trabajar a tiempo completo en el mismo” (p. 37). A su vez, se propone que el equipo sea contratado directamente por el banco para evitar costos asociados a la consultoría de perfiles que brinden servicios de manera externa.

En este marco, los principales roles necesarios previstos para la implementación serán los siguientes:

Project Manager: es el responsable de la planificación, coordinación y seguimiento de todas las actividades del proyecto. Supervisa el cumplimiento de los plazos, la asignación de recursos y la gestión de riesgos. Además, actúa como enlace entre la dirección del banco y el equipo técnico, asegurando que las decisiones se tomen con base en criterios de valor y factibilidad. Según Project Management Institute (2013), es “la persona nombrada por la organización ejecutante para liderar al equipo que es responsable de alcanzar los objetivos del proyecto” (p. 540).

Arquitecto Salesforce: define la estructura técnica general de la solución. Diseña la arquitectura del sistema y establece los flujos de integración entre el asistente virtual, Data Cloud y Service Cloud. Su experiencia en la plataforma Salesforce permite anticipar limitaciones técnicas y seleccionar los componentes adecuados para lograr escalabilidad y seguridad, así como también tomar decisiones sobre las posibles alternativas de implementación que permite la flexibilidad de la plataforma.

Consultor Funcional Senior Salesforce: traduce los requerimientos del negocio en configuraciones concretas dentro de la plataforma. Colabora con las áreas de Atención al Cliente y Tecnología para modelar los flujos de interacción que el agente ejecutará. Este perfil

asegura que las funcionalidades desarrolladas respondan a las necesidades reales del banco y cumplan con las políticas de operación vigentes.

Consultor Funcional Salesforce: acompaña la implementación del agente con las primeras configuraciones y activaciones en la plataforma, diseño de formatos de página en el sistema, y la implementación de campos con tipos de datos y acceso a los mismos según se haya definido en el alcance.

Desarrollador de código Apex: implementa personalizaciones necesarias en el entorno Agentforce que requieran un nivel alto de complejidad, muchas veces cubiertas con automatizaciones de flujos (Flows) o creación de componentes con código Apex como clases o triggers. También participa en la construcción de prompts y pruebas sobre la plataforma para asegurar la operatividad técnica.

Data Engineer: gestiona la calidad, disponibilidad y estructura de los datos utilizados por el agente. Se encarga de definir las fuentes de datos que deberán ser migrados a Salesforce y preparar los conjuntos de datos. También colabora en la detección de sesgos o errores de interpretación en los resultados del modelo.

Analista de QA (Quality Assurance): perfil encargado de validar el correcto funcionamiento del agente en todas las etapas del desarrollo. Ejecuta pruebas funcionales y de seguridad en el entorno sandbox (ambiente de desarrollo y pruebas), verificando la precisión de las respuestas, la integridad de los datos y la estabilidad del sistema antes de su despliegue a un ambiente productivo.

Además de los perfiles mencionados como necesarios, se propone también de manera opcional la inclusión de un Product Owner, un perfil que deberá contar con un gran conocimiento del organigrama y ecosistema general del banco que brindará asesoría general sobre las rutas de desarrollo que serán posibles en el banco en general. Habitualmente es un perfil interno proporcionado por cada banco que no está asignado como parte de trabajo del equipo, sino que participa en ciertas reuniones claves para acompañar con decisiones para asegurar que las decisiones técnicas se tomen de manera alineada con la planificación estratégica del banco como organización.

Cada uno de los perfiles contribuye a un aspecto crítico del proyecto. La interacción entre los distintos roles se articula bajo una metodología ágil, con reuniones de seguimiento semanales y una estructura jerárquica clara que permite escalar decisiones rápidamente. La metodología de trabajo ágil se consolida tomando las herramientas más valiosas del framework conocido como Scrum, la cual según Martel (2014), es “un conjunto de buenas prácticas para

la gestión de proyectos” que propone realizar entregas parciales y regulares del producto final (p. 12). En la propuesta el asistente virtual funcional será el producto final que podrá ser implementado primero con una serie de casos de uso y luego, continuado en futuras implementaciones, podrá recibir mejoras evolutivas. Cada lote de trabajo en el que se agrega valor incremental a la implementación será un Sprint de trabajo en el que el equipo de trabajo deberá notificar los avances. Para lograr esto, los líderes del proyecto acordarán una lista de requerimientos simplificada con las tareas que el equipo pueda hacer durante las próximas 2 semanas, de acuerdo con el cronograma general de actividades. A ese periodo de 2 semanas se lo conoce como Sprint porque durante ese tiempo se realizan las tareas necesarias para cumplir los objetivos de corto plazo y probar todas las funcionalidades que el equipo se comprometió a terminar en ese período. La duración habitual de un Sprint debe definirse con una duración de entre 1 y 4 semanas (Martel, 2014, p. 14). En el caso de esta propuesta se propone mantener Sprints de trabajo con una duración de 2 semanas, una extensión acorde al tipo de proyecto de implementación de una plataforma cloud, a pesar de que esto mismo podrá ser validado y revisado con el banco durante la primera etapa de Diseño.

4.4.2 Recursos Tecnológicos

Los recursos tecnológicos representan la base técnica sobre la cual se desarrollará el asistente virtual inteligente del banco. La solución se implementará utilizando la infraestructura de Salesforce, la cual funciona como una plataforma accesible de manera global con un plan de suscripción anual. A continuación, se indicarán los recursos de licenciamiento y materiales que serán requeridos para la puesta en marcha del asistente virtual:

Licencias de Salesforce. La cantidad de usuarios nominales a adquirir dependerá de la cantidad de personas que desea incluir el banco como usuarios de gestión back-end de la solución. Para la implementación del modelo de agente sólo se requiere una licencia activa con los módulos o nubes de Data Cloud, Service Cloud y Agentforce. Para realizar una estimación de costos y calcular lo necesario al momento de comenzar el proyecto, se incluyó un total de 10 licencias nominales que cubra todas las especificaciones de producto mencionadas.

Licencias para consumo de créditos de LLM. La principal funcionalidad será el asistente virtual, por lo que se deberá adquirir un paquete de créditos que permita la ejecución del motor de razonamiento que vincula el desarrollo con el LLM utilizado. El costo del procesamiento de GPT será incluido en este costo como parte del licenciamiento que Salesforce denomina Flex Credits y que asegura una conexión completa y transparente con el modelo de lenguaje de gran escala sin requerir ningún otro sistema intermedio.

Ambientes de desarrollo, pruebas y productivo. La implementación requiere como mínimo que las licencias mencionadas anteriormente sean activadas en una instancia con licenciamiento de Salesforce Enterprise, lo que garantiza la posibilidad de crear ambientes de pruebas y desarrollo.

Computadora local con instalación de Data Loader. La herramienta que será utilizada para la migración de datos debe ser disponibilidad en un equipo con Windows o Mac que ponga a disposición el banco para asegurar un entorno seguro en el tratamiento de datos a migrar a Salesforce. Se recomienda la utilización de un equipo con capacidad de procesamiento por encima de un procesador Intel i7 o su equivalente AMD Ryzen 7 en Windows y un procesador M1 para computadoras Mac. El dispositivo debe contar con acceso a Internet de manera obligatoria. A efectos de compartir los costos mínimos necesarios, se consideró una CPU que sea capaz de gestionar grandes volúmenes de registros y reducir significativamente los tiempos de ejecución durante las operaciones de carga y verificación de datos. En escenarios de carga intensiva o en procesos de migración que implican millones de registros (como ocurre habitualmente en bancos con cientos de clientes), un procesador de esta gama se considera esencial para mantener un rendimiento sostenido. La memoria RAM constituye otro factor crítico en este tipo de tareas, ya que Data Loader se ejecuta sobre una máquina virtual de Java (JVM) y utiliza la memoria disponible para almacenar temporalmente los datos procesados. Se recomienda disponer de 32 GB para trabajar con conjuntos de datos extensos o con archivos CSV de gran tamaño que contienen decenas de miles de filas. En cuanto al almacenamiento, se sugiere el uso de una unidad de estado sólido (SSD) tipo NVMe de 1 TB de capacidad, con la recomendación de contar con al menos 100 GB de espacio libre para garantizar que el sistema pueda manejar los archivos de entrada y los temporales que se generan durante la ejecución del proceso.

Software. La computadora mencionada en el punto anterior deberá contar con un sistema operativo compatible con la versión más actualizada de Data Loader al momento de comenzar el proyecto. La estimación de costos ha sido planificada con una licencia de Windows 11 (64 bits). Además, es imprescindible contar con una instalación del Java Runtime Environment (JRE) versión 11 o superior, configurar correctamente la variable de entorno JAVA_HOME para garantizar el funcionamiento de la aplicación, y utilizar archivos de entrada con los datos a migrar entregados por el banco en formato CSV codificado en UTF-8 y libres de fórmulas o caracteres especiales. También se debe contar en el dispositivo con una licencia

de Microsoft 365 para Empresas, la cual permite utilizar sin restricciones la herramienta Microsoft Excel para gestionar los archivos de importación.

4.5 Costos del Proyecto

El desarrollo e implementación del asistente virtual implicará una inversión inicial destinada a cubrir los costos de licencias, y servicios profesionales. A diferencia de otros proyectos de software tradicionales, el modelo de costos asociado a la plataforma Salesforce se basa en un esquema de suscripción y consumo, lo que permite al banco administrar los gastos de acuerdo con el uso real del sistema. Para facilitar el ejercicio se utilizará la moneda dólar estadounidense como referencia para los cálculos del personal que debe formar parte del equipo de trabajo y de los componentes técnicos que se requieren adquirir.

4.5.1 Costos del Equipo de Trabajo

El equipo de trabajo estimado se corresponde con valores del mercado en octubre de 2025, considerando los salarios aproximados de los perfiles incluidos con residencia dentro del territorio argentino.

Tabla 3

Costos de Perfiles Asignados al Proyecto

Perfil	Salario Mensual (USD)
Project Manager	3.200
Analista de Negocios	2.000
Tech Lead	4.500
Arquitecto Salesforce	3.000
Consultor Funcional Senior Salesforce	2.000
Consultor Funcional Salesforce	1.500
Desarrollador Salesforce	2.800
Data Engineer	1.800
Analista de QA	1.400
Costo Mensual	22.200

Nota. Elaboración propia.

4.5.2 Costos Técnicos

Para simular el ejercicio de costos se definió como medida un banco con una cartera de 5 millones de clientes, de los cuáles un 10% se comunica todos los meses por canales digitales en los que funcionará el asistente virtual, es decir, unos 500.000 clientes activos con consultas mensuales. La estimación de los costos de Data Cloud fue realizada de acuerdo con la Calculadora de Precios que Salesforce ofrece de manera gratuita para conocer un aproximado de consumo (Salesforce, 2025f). Para lograr el dimensionamiento, la calculadora ofrece la configuración de una serie de parámetros que fueron adaptados al contexto mencionado anteriormente.

Tabla 4

Simulación de Parámetros Requeridos para Data Cloud

Parámetro	Estimado	Descripción
Perfiles de cliente no único integrados	500.000	Clientes activos que requieren consultas
Total de perfiles	750.000	Se incrementan los registros en 50% por nuevos clientes
Porcentaje de datos actualizados a diario	10%	Movimientos, reclamos, cambios de contacto
Datos no estructurados (documentos)	Si	Documentos internos del banco
Cantidad de registros referenciados a través de Zero Copy	2.000.000	Operaciones históricas y logs sin replicar
Porcentaje de datos en streaming	5%	Eventos en tiempo (transacciones/reclamos)
Datos en tiempo real en fracciones de segundos	No	Latencia de segundos/minutos es suficiente
Perfiles unificados entre fuentes	Si	Perfil único de cliente
Actualización diaria de las métricas clave	25	Atención, SLA, NPS, uso del asistente
Consultas totales diarias de un máximo de 10 usuarios	300	Analistas, supervisores
Búsquedas/análisis de datos no estructurados por día	200	PDF, conversaciones
Segmentos publicados por año	50	Productos, comportamiento
Activaciones de segmentos por año	100	Campañas internas
Porcentaje de datos compartidos a bases de datos externas	5%	Reportes regulatorios, sistemas corporativos

Nota. Elaboración propia de acuerdo con la Calculadora de precios, por Salesforce, 2025f.

Con la configuración presentada se ejecutó la calculadora de precios, brindando un resultado final de 17.866 dólares anuales con la contratación de Data Cloud bajo la modalidad de suscripción (Salesforce, 2025f, p. 3).

Figura 20

Presupuesto para la contratación de Data Cloud (expresado en dólares)



Nota. Obtenido de *Data Cloud pricing calculator* utilizando parámetros ingresados de manera personalizada para esta propuesta, por Salesforce, 2025f.

Agentforce también debe ser abonado como una licencia más, dado que consume créditos en el funcionamiento del motor de razonamiento con el LLM. La plataforma cuenta con un licenciamiento que permite contratar paquetes de créditos o por el contrario directamente abonar por cada conversación que el asistente comercial inicie con un cliente. Dado que permite mayor flexibilidad y control, se recomienda la adquisición de licencias de paquetes de créditos llamadas Flex Credits, las cuales incluyen 100.000 créditos de LLM por cada 500 dólares cobrados. Para comprender el real alcance de estos límites se debe considerar que una acción del asistente virtual, como por ejemplo crear un reclamo de un cliente, consume

unos 10 centavos de dólar por cada vez que se ejecuta, lo que se corresponde con el consumo de 20 créditos para ambientes productivos y 16 créditos (8 centavos de dólar) para ambientes de desarrollo o pruebas. Además, Salesforce impone un límite de 10.000 tokens que pueden ser utilizados en la comunicación con el LLM, equivalente aproximadamente a 40.000 caracteres, 8.000 palabras o 22 hojas de un documento de hoja tipo A4. La relación entonces indica que aproximadamente el motor de razonamiento consume un token por cada 4 caracteres que sean utilizados. Para el ejercicio de estimación se estima la compra de 3 paquetes mensuales, es decir, 1.500 dólares mensuales que habilitarán un total de 300.000 créditos mensuales. Salesforce requiere la compra de al menos un año de licenciamiento, por lo que el costo anual de Agentforce será de 18.000 dólares.

Figura 21

Licenciamiento Recomendado de Agentforce



Nota. Obtenido de *Agentforce Pricing*, por Salesforce, 2025c.

Con respecto a la computadora que se requiere para los trabajos de migración, se recomienda que las personas asignadas al proyecto puedan trabajar con sus dispositivos propios o con un dispositivo ofrecido por el banco por el período que dure el proyecto. Sin embargo, a modo de considerar los posibles costos, se incluye la cotización de una computadora de la marca Lenovo en su línea de equipos para trabajar. Se propone la compra de la computadora

ThinkPad E14 Gen 7 (14" Intel), la cual cuenta con un procesador Intel Core 7, 32 GB de memoria de tipo DDR5-5600MT/s, una unidad de disco primaria de 512 GB en formato SSD M.2 y una pantalla de 14 pulgadas con IPS. El sistema operativo utilizado será Windows.

Figura 22

Computadora Recomendada para el Proceso de Migración de Datos



6 CUOTAS SIN INTERÉS

ThinkPad E14 Gen 7 (14" Intel)

¡Personalizable!

★★★★★ 5.0 (6)

Precio Web \$2.941.582,84

\$2.235.603,20 23% de descuento IVAInc.

Precio sin impuestos nacionales: \$2.023.170,24

Arma tu PC

☐ Comparar

Nota. Obtenido de *Especificaciones y precios en Argentina*, por Lenovo, 2025.

Dado que el valor de compra en Argentina para la computadora Lenovo es de 2.235.603,20 pesos, en una conversión de 1.475 pesos argentinos por cada dólar estadounidense, se estima el valor de la computadora en 1.516 dólares.

La licencia de Windows, Data Loader y Java Runtime Environment, identificados como otros componentes de software necesario para la ejecución del proyecto se encuentran, ya se encuentran estimados dentro del presupuesto de la computadora, dado que el sistema operativo ya estará instalado en la máquina y las otras dos herramientas mencionadas no requieren el pago de licenciamiento. Con respecto a la licencia de Microsoft Office necesaria para preparar los archivos que serán migrados al sistema, queda a definición del banco la posibilidad de incluir a las personas del equipo de trabajo como parte de usuarios de la Suite corporativa o proveer de una herramienta similar que cumpla con la capacidad de crear y modificar archivos de tipo CSV.

Luego de identificar cada ítem necesario, se presenta la cotización final con los recursos humanos y técnicos necesarios para ejecutar el proyecto con un costo de 126.182 dólares. En

el ejercicio no se considera un margen de ganancias sobre el equipo de trabajo porque la propuesta presenta la posibilidad de que el banco contrate directamente a un equipo interno para el desarrollo del asistente virtual, equipo que podrá continuar trabajando en el banco luego de finalizar el proyecto. En caso de requerir un proveedor externo que realice el trabajo, se deberá realizar una nueva estimación que incluye los posibles márgenes de ganancias.

Tabla 5

Costo Total del Proyecto

Concepto	Costo Estimado (USD)	Descripción
Recursos Humanos	88.800	22.200 USD del equipo asignado por 4 meses de trabajo
Licenciamiento	35.866	18.000 USD más 17.866 USD
Equipos de trabajo	1.516	Computadora de única compra
Costo Final	126.182	

Nota. Elaboración propia.

4.6 Cronograma de tareas

La implementación del asistente virtual inteligente se planificará en un cronograma de trabajo que contempla distintas fases interdependientes. Cada fase representa un avance progresivo en la madurez del sistema, desde la definición de los requerimientos hasta el despliegue y seguimiento de los resultados. Esta estructura temporal tiene como objetivo asegurar la calidad técnica del desarrollo y la coherencia funcional con los procesos del banco, evitando interrupciones en los servicios existentes.

El proyecto iniciará con una etapa de planificación y relevamiento, en la que se presentará a cada banco el alcance propuesto del asistente virtual, los objetivos operativos y los indicadores de éxito que permitirán evaluar su impacto. La documentación generada servirá como base para las fases posteriores.

Una vez definido el marco de trabajo, se dará comienzo al desarrollo funcional y técnico dentro del entorno Salesforce. El equipo técnico llevará adelante las pruebas unitarias y

funcionales necesarias para verificar la correcta ejecución de las acciones y la precisión de las respuestas generadas. Durante este proceso, se mantendrá una comunicación constante entre los consultores funcionales y los desarrolladores, a fin de asegurar la alineación entre la lógica del negocio y la configuración técnica del agente.

Posteriormente se realizará una fase de validación integral, que incluirá la revisión de los resultados de las pruebas, el ajuste de configuraciones y la evaluación del desempeño del asistente en un entorno de prueba (sandbox). Este periodo será supervisado por los equipos de calidad, analistas de datos y responsables de cumplimiento, quienes verificarán que el sistema opere dentro de los parámetros de seguridad, exactitud y confiabilidad definidos por la organización.

Una vez alcanzados los niveles esperados de confiabilidad y aceptación, el proyecto avanzará hacia la fase de estabilización. Esta etapa incluirá un acompañamiento al banco en las tareas diarias que realice el asistente virtual, a efectos de evitar cualquier incumplimiento del alcance acordado para el proyecto.

El cronograma completo de trabajo está previsto para desarrollarse en un período aproximado de cuatro meses, con una primera fase de planificación, refinamiento de procesos y diseño con una duración de 5 semanas, seguida por una fase de implementación, pruebas, despliegue y estabilización que se extenderá durante el resto del ciclo. Para la construcción de esta planificación se estimaron una serie de actividades técnicas que deben cumplirse con la cantidad de días especificados para cada una de ellas y con una determinada participación para cada perfil asignado al proyecto. En el siguiente cronograma se observa la planificación en semanas.

Tabla 6

Etapas del Proyecto en Semanas

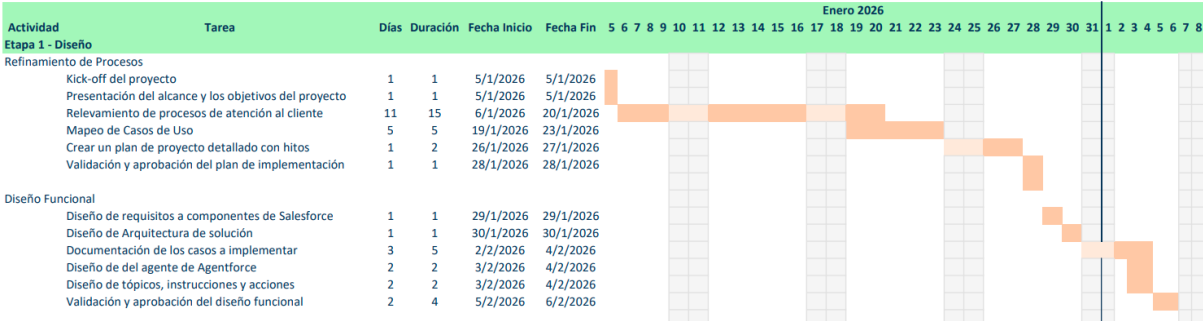
Actividades	Enero 2026				Febrero 2026				Marzo 2026				Abril 2026			
	Semana 1	Semana 2	Semana 3	Semana 4	Semana 1	Semana 2	Semana 3	Semana 4	Semana 1	Semana 2	Semana 3	Semana 4	Semana 1	Semana 2	Semana 3	Semana 4
Etapas																
Etapas 1 - Diseño																
Refinamiento de Procesos																
Diseño Funcional																
Etapas 2 - Implementación																
Configuración Inicial																
Desarrollo Técnico																
Pruebas																
Despliegue a Producción																
Estabilización																

Nota. Elaboración propia del Diagrama de Gantt con Microsoft Excel.

Como convención para la presentación del cronograma se definió el lunes 5 de enero de 2026 como el día de comienzo del proyecto, tal como se presentará en las imágenes del

diagrama de Gantt del proyecto. Para la primera fase del proyecto, se propone las actividades de Refinamiento de Procesos y Diseño Funcional con su duración de 5 semanas.

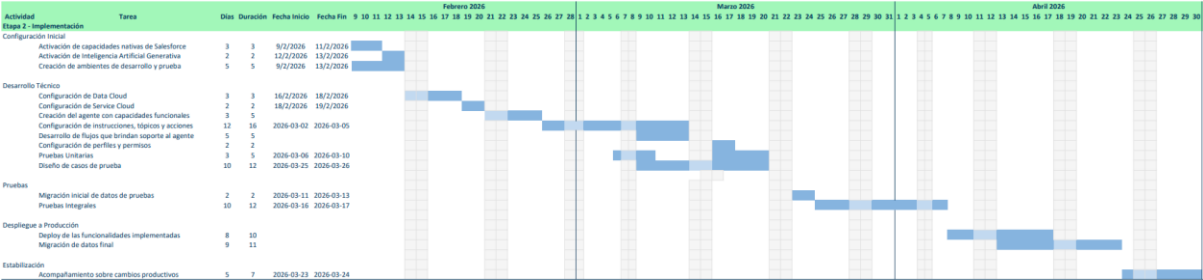
Tabla 7
Actividades de la Etapa 1 del Proyecto



Nota. Elaboración propia del Diagrama de Gantt con Microsoft Excel.

La segunda etapa contendrá las actividades para realizar la Configuración Inicial, el Desarrollo Técnico, las Pruebas, el Despliegue a Producción y Estabilización. Para una mejor visualización, se presenta a continuación cada una de las actividades de manera consecutiva con sus respectivos hitos que permitirán controlar el avance el proyecto.

Tabla 8
Actividades de la Etapa 2 del Proyecto



Nota. Elaboración propia del Diagrama de Gantt con Microsoft Excel.

4.7 Métricas de Desempeño y Mejora Continua

La evaluación del desempeño del asistente virtual constituye una parte esencial para garantizar la adopción de la nueva tecnología, ya que permite medir su impacto, controlar la calidad de sus respuestas y garantizar que su funcionamiento se mantenga alineado con los objetivos del banco. Según Salesforce (2025a), los proyectos de Agentforce deben incorporar

desde su etapa de diseño un sistema de monitoreo continuo que permita obtener información en tiempo real sobre la efectividad del agente, su precisión y el grado de satisfacción del cliente.

Las métricas principales se concentrarán en tres dimensiones: operativa, de calidad de servicio y de valor estratégico. En el plano operativo, los bancos que implementen el sistema deberán analizar el volumen de interacciones atendidas por el asistente y la proporción de consultas resueltas sin intervención humana. Este indicador, conocido como tasa de resolución automática, refleja la eficiencia del agente y su capacidad para reducir la carga de trabajo del personal de atención al cliente.

La dimensión de calidad de servicio evaluará la precisión de las respuestas y el nivel de satisfacción de los usuarios. Este último se podrá ser medido a través de encuestas automáticas y del Índice Neto de Promotores (NPS, por sus siglas en inglés), que permite identificar la percepción del cliente respecto de la utilidad y facilidad de uso del asistente. Según Gotts et al. (2024), toda implementación de agentes conlleva monitorear la eficacia del agente y la transferencia de datos a un humano, lo que ayuda a definir luego qué capacidades adicionales se podrán añadir para evolucionar una primera implementación (p. 30).

En la dimensión estratégica, Gotts et al. (2024) proponen plantear implementaciones con proyectos que sean específicos para obtener resultados de manera inmediata adoptando un enfoque de cambio de procesos, gobernanza de datos y gestión de metadatos, lo que significa que se debe considerar documentar todos los procesos de manera interna en el banco, consolidar los datos adecuados para proveer al agente y poner a disposición los metadatos con la cual los agentes aprender el negocio bancario (p. 31). Luego, las métricas que deberá analizar el banco se enfocarán en los beneficios económicos y organizacionales derivados del uso del asistente, asumiendo que la automatización de las consultas frecuentes permitirá disminuir los costos relacionados con tareas repetitivas y optimizar la asignación de recursos humanos hacia funciones de mayor valor agregado. Además, la disponibilidad de atención digital continua reforzará la competitividad de la institución frente a otras entidades del sector que aún dependen de canales más tradicionales.

El proceso de mejora continua estará basado en un modelo iterativo de aprendizaje y ajuste. Las métricas recolectadas se podrán revisar periódicamente en reuniones entre los equipos técnicos, analistas de datos y responsables de negocio, con el objetivo de identificar oportunidades de optimización. Cada ciclo de mejora debe incluir el control de consumo de tokens, la actualización de prompts que utiliza el asistente virtual, la revisión de los tópicos del

agente y la incorporación de nuevos flujos funcionales en función de las necesidades del cliente o del negocio.

4.8 Factores condicionantes

La implementación del asistente virtual inteligente en el sector bancario está influida por un conjunto de factores internos y externos que determinan su viabilidad y alcance. Estos factores se dividen principalmente en dimensiones regulatorias, tecnológicas y socioculturales, e inciden de manera directa en las estrategias de diseño, desarrollo y gestión del proyecto. La comprensión de este entorno es esencial para anticipar riesgos y asegurar una adopción responsable de la inteligencia artificial en un contexto altamente regulado como el financiero.

4.8.1 Factores internos

Factor humano. La disponibilidad de tiempo, la dedicación de los equipos de negocio y la existencia de liderazgo interno de los bancos tendrá influencia directa en la adopción de la solución. Los proyectos basados en inteligencia artificial requieren la participación de los distintos sectores de la organización, y en este caso especialmente de las áreas de atención al cliente, tecnología y cumplimiento normativo. El compromiso de estos equipos garantizará que los requerimientos funcionales del asistente respondan a las necesidades reales del banco y de sus clientes. Además, es indispensable que la institución cuente con líderes internos que actúen como promotores del proyecto, capaces de impulsar la iniciativa, tomar decisiones oportunas y sostener la colaboración entre los equipos técnicos y operativos.

Acceso a la información. El banco deberá asimismo garantizar el acceso a los datos de los clientes para la migración inicial propuesta en el proyecto, lo cual podría requerir un tratamiento de datos con procesos de normalización o limpieza de datos, lo que se encuentra por fuera del alcance del proyecto. Los responsables de negocio dentro del banco son aquellas personas que pueden garantizar que se realicen los esfuerzos necesarios para que los datos utilizados sean adecuados, ya que, como todo sistema automatizado con aplicación de inteligencia artificial, los asistentes virtuales requieren que los datos que son provistos tanto para el entendimiento del contexto como para conocer las preferencias de los clientes sean correctos y actualizados.

Contexto tecnológico del banco. La madurez de los sistemas bancarios y su capacidad de integración determinan en gran medida el éxito de la automatización. La ausencia de una estructura tecnológica coherente podría limitar la capacidad del agente para acceder a datos confiables en tiempo real o ejecutar acciones de manera automatizada.

4.8.2 Factores externos

Protección de datos personales. En el plano regulatorio, la protección de los datos personales constituye una condición determinante. La Ley 25.326 establece obligaciones precisas sobre la obtención, almacenamiento y tratamiento de la información sensible de los usuarios. Para cumplir con esta normativa, las entidades financieras deben garantizar que los datos sean utilizados únicamente con fines legítimos, manteniendo altos niveles de seguridad y confidencialidad.

Grado de aceptación del asistente virtual por parte de los usuarios. Si bien la digitalización de los servicios financieros en Argentina ha crecido de forma sostenida, todavía existen segmentos de la población que muestran cierta resistencia a los sistemas automatizados, principalmente por motivos de confianza o preferencia por la atención humana. La percepción positiva del cliente dependerá en gran medida de la transparencia con la que el banco comunique las funciones del asistente y de la calidad de la experiencia de uso que este proporcione. En este sentido, la interacción debe ser empática y coherente con los valores del banco, garantizando siempre la posibilidad de acceder a asistencia humana en consultas complejas.

Constante control de consumo. La solución propuesta ofrece la gran ventaja de no requerir infraestructura local para que el asistente virtual sea funcional, pero debido a que será implementado bajo una modalidad de suscripción de licencias de Salesforce, el banco deberá monitorear activamente el consumo de tokens que requiere el agente para evitar consumos desproporcionados por eventos específicos o posibles problemas de seguridad que afecten al banco en general.

Capítulo 5 – Prototipo de Asistente Virtual

5.1 Introducción

Con el propósito de validar la factibilidad técnica de la propuesta presentada en el capítulo anterior, se diseñó un prototipo funcional del asistente virtual utilizando la plataforma Salesforce. Este prototipo no constituye una implementación final del proyecto, sino una demostración experimental desarrollada para evidenciar de manera práctica cómo se comporta un agente inteligente dentro del entorno de Salesforce y cómo podría integrarse con los procesos de atención al cliente de una entidad bancaria. El objetivo principal de esta prueba es verificar la operatividad básica del modelo en un entorno controlado, explorando sus componentes fundamentales: la creación de tópicos y de instrucciones, la configuración de acciones personalizadas para la ejecución de tareas simples, la interacción con los módulos Service Cloud y Data Cloud y la conexión con sistemas externos que podrían funcionar como bases de datos o punto de acceso a los datos.

La implementación se realizó en un entorno de Desarrollo (Developer) gratuito de Salesforce, lo que permitió experimentar con la configuración del agente sin comprometer información real ni afectar sistemas productivos. Según Salesforce Help (2025b), este entorno reproduce las condiciones técnicas de una organización de trabajo contratada por una empresa y ofrece las herramientas necesarias para el desarrollo, prueba y supervisión de agentes virtuales de forma gratuita, ya que se puede solicitar en línea sin costo con ciertas limitaciones técnicas que Salesforce aplica a la instancia de trabajo. La principal limitante es que sólo se permite acceso a dos usuarios y el espacio de almacenamiento de datos en base de datos se limita a 5 MB y a 20 MB para archivos. Considerando que un registro como la ficha de datos principales de un cliente puede ocupar unos 2 KB de almacenamiento, sólo se podrán crear un máximo aproximado de 2.560 registros. En relación con posibles integraciones con sistemas externos, la plataforma permite aproximadamente 5.000 llamadas API cada 24 horas. Con respecto a los procesos, las pruebas del piloto incluyen consultas frecuentes de atención al cliente con respuestas generadas a partir del procesamiento de lenguaje natural y con utilización de RAG, y la creación de reclamos simulados que serán persistidos en Service Cloud, la interfaz gráfica que representa el acceso de usuarios internos del banco.

La realización de este prototipo presenta la viabilidad técnica de la propuesta, demostrando que los componentes teóricos más destacados descritos en el Capítulo 4 pueden configurarse efectivamente dentro de la plataforma a partir de una aproximación empírica que

sirve como evidencia del potencial de Agentforce para adaptarse a los flujos de negocio particulares de cada institución financiera.

5.2 Configuración del Entorno de Prueba

El acceso a la instancia de trabajo se obtuvo a través del registro de una solicitud en el portal de desarrolladores de Salesforce, lo cual crea una organización de trabajo con todas las funcionalidades principales de la plataforma, incluyendo una interfaz de usuario para configurar el agente, solicitudes de API y capacidades de personalización mediante configuración, herramientas declarativas y desarrollo de código Apex, el lenguaje de programación que utiliza Salesforce con tipado estricto para ejecutar instrucciones de control de flujo y transacciones en los servidores donde residen las instancias de trabajo (Salesforce, 2024a, p. 1).

La edición Developer se utiliza únicamente con propósitos de aprendizaje, pruebas o prototipado. Utilizar entornos aislados para desarrolladores evita interferir con sistemas compartidos. Una vez activada la organización de trabajo, se procedió a habilitar las funciones de Einstein y Agentforce necesarias para construir el asistente virtual. Esto se realizó desde el acceso a la Configuración con un usuario con perfil de administrador. Luego, en Agentforce Studio se creó un nuevo agente mediante la interfaz de Agentforce Builder, donde se diseñan y gestionan los agentes de IA. Adicionalmente, se configuraron componentes de Service Cloud relevantes para persistir datos y se habilitó una base de conocimiento interna, mediante el asistente de Service Setup. Esto permitió cargar artículos con los documentos PDF que fueron creados para simular manuales de una base de conocimiento del banco.

Para brindarle mayor realismo a la implementación del prototipo, se definió Banco Horizonte como una organización bancaria ficticia, que permite representar en la construcción del asistente virtual a un supuesto banco argentino con clientes para el que opera el asistente virtual. El nombre fue elegido con fines prácticos para brindarle información sobre el banco ficticio al asistente virtual y comprobar que puede comprender el contexto en el que se desarrolla. A su vez, los archivos de base de conocimiento generados fueron contruidos de manera manual, incluyendo detalles ficticios sobre canales de contacto, información del banco y posibles procedimientos internos, simulando la operatoria de un banco real.

5.3 Desarrollo Inicial del Agente

5.3.1 Configuración Básica

Para demostrar la viabilidad de la implementación con algunos casos de uso limitados, se planificó la configuración de ciertos flujos de acción y personalizaciones diseñadas en código Apex. Luego de configurar los parámetros iniciales del asistente virtual, se procedió a crear un primer tópico de saludo inicial para que el asistente virtual comprenda cómo saludar a un cliente. Se configuró también un tópico de validación de cliente para solicitarle datos al cliente cuando quiere manipular datos personales o solicitar un reclamo, y se diseñaron los archivos en formato PDF para proveerle al asistente de información básica del banco. Finalmente, se configuró una acción para crear reclamos de manera automatizada, los cuales podrán ser gestionados más tarde por equipos conformados por asistentes humanos del banco.

5.3.1 Configuración del Asistente Virtual

Desde un ambiente seguro al que se accede sólo con usuario y contraseña, se configuró al asistente virtual con el nombre de Hori, proveniente de las primeras cuatro letras del nombre Horizonte, con el objetivo de simular interacciones con clientes. Salesforce provee una consola integrada en la plataforma para crear y personalizar agentes de IA de forma declarativa. Desde esta interfaz, se creó el agente basado en la plantilla de *Service Agent* (Agente de Servicio) y se personalizó su configuración según los requerimientos bancarios desarrollados en la propuesta del prototipo. Agentforce Studio incluye el módulo Agent Builder, que a su vez ofrece utilidades como el *Prompt Builder* (constructor de instrucciones) para definir las instrucciones que se utilizan para guiar al asistente virtual en su ejecución, y la gestión de tópicos y acciones que conforman su comportamiento. A continuación, se detallan los principales aspectos desarrollados para Hori dentro de esta instancia de trabajo en la nube:

Figura 23

Configuración Inicial en Entorno de Desarrollo

* Name	Created On
Agentforce Bank Service Agent	October 17, 2025 at 05:15 PM
* API Name	Created By
Agentforce_Bank_Service_Agent	Carlos Manuel Lopez
* Description ⓘ	Last Modified By
Ofrece interacciones personalizadas a clientes en español mediante un agente autónomo de IA. El agente de servicio bancario brinda asistencia inteligente a sus clientes con consultas comunes y deriva los casos complejos relacionados con Banco Horizonte.	Carlos Manuel Lopez
* Role ⓘ	
Un agente de atención al cliente bancario con inteligencia artificial cuya función es ayudar a los clientes con preguntas y ayuda en general en español.	
* Company ⓘ	
Su empresa, Banco Horizonte, se especializa en brindar servicios financieros a clientes particulares argentinos. Ofrece productos como cuentas corrientes, cuentas de ahorro, tarjetas de débito y crédito, así como préstamos con garantía e hipotecarios.	

Nota. Obtenido de la instancia de trabajo de Salesforce con elaboración propia de las configuraciones.

En la creación de Hori, se utilizó un prompt global que permite establecer su rol y tono. Mediante el Prompt Builder se redactaron instrucciones descriptivas en lenguaje natural indicando que Hori es un asistente virtual bancario diseñado para ayudar a clientes del Banco Horizonte, con consultas sobre productos y servicios financieros. La denominación del Banco Horizonte y todos los datos proporcionados al asistente virtual fueron simulados para representar a un banco que permita demostrar la ejecución del agente, sin involucrar datos de bancos o clientes reales en este proceso de prueba.

Las instrucciones globales funcionan como directrices que el agente siempre seguirá en cualquier conversación, como por ejemplo siempre saludar cortésmente, usar terminología apropiada del banco o evitar compartir o solicitar información confidencial cuando no sea necesario. Salesforce permite especificar estas instrucciones globales para alinear el comportamiento del agente e incluso aprovechar las capacidades de IA generativa para sugerir la configuración inicial del agente a partir de dichas descripciones. En este caso de prueba, se incluyeron indicaciones sobre límites que le indiquen al asistente virtual que no brinde asesoramiento financiero personalizado ni respuestas fuera de su base de conocimiento y políticas de escalamiento a personas. Además, se aprovecharon las barreras de seguridad

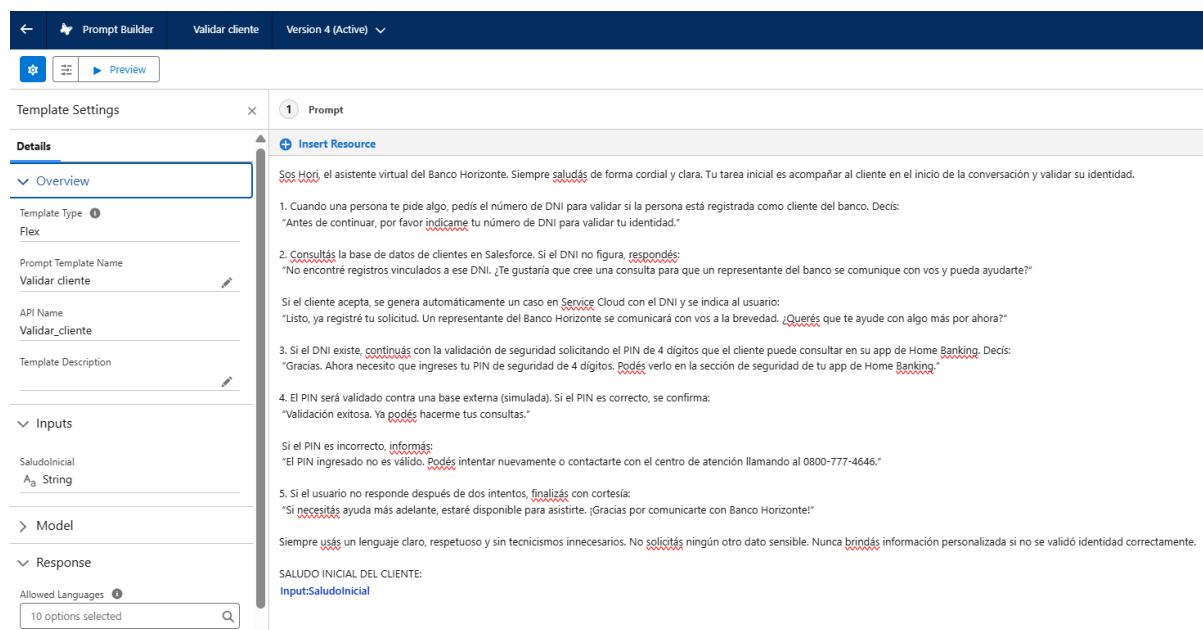
proporcionadas por Agentforce en su conexión con el LLM utilizado (GPT), para evitar que el asistente se desvíe de sus instrucciones principales o genere contenido inapropiado, bloqueando conversaciones fuera de tópico o potencialmente indebidas, lo que permite mitigar riesgos de respuestas alucinadas o sesgadas. Es decir que siempre que el agente no pueda responder algo por falta de información o de comprensión, va a sugerir que el usuario vuelva a comunicar su pedido de otra manera. En caso de que continúe la falta de entendimiento, el asistente virtual cuenta con la posibilidad de derivar la atención a una persona, generando automáticamente una consulta o reclamo que será dirigido al sistema interno del banco para su posterior gestión fuera de línea. Estas configuraciones iniciales garantizan que Hori mantenga un tono consistente y confiable acorde a las políticas del banco durante toda la interacción con un cliente.

5.3.2 Configuración de Acciones Personalizadas

Para dotar al asistente virtual de capacidades operativas, se configuraron acciones asociadas a los tópicos. Las acciones son funciones que el agente puede ejecutar para cumplir una solicitud, ya sea operaciones estándar (consultar datos, resumir registros o automatizar el envío de un correo electrónico) o personalizadas que encapsulan lógica de negocio más compleja. En el prototipo se configuró una acción principal como punto de entrada para las conversaciones para que el asistente virtual pueda validar la identidad de los clientes antes de realizar cualquier gestión. Para ello se configuró el tópico “Validación de cliente” que contiene una acción creada a partir del prompt “Validar cliente” con las instrucciones de solicitud de información para realizar la validación de identidad del cliente. En la siguiente imagen se presenta la configuración final de la instrucción que se utiliza para guiar al asistente virtual a ejecutar acciones.

Figura 24

Prompt para Validar Cliente al Iniciar una Conversación

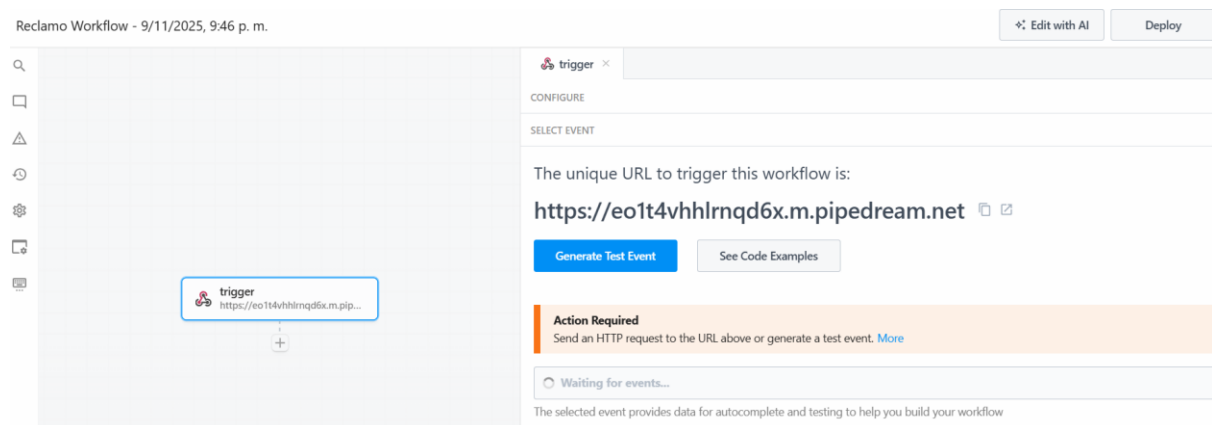


Nota. Elaboración propia en la configuración de Prompt Builder de Salesforce.

Adicionalmente, se implementaron acciones personalizadas para extender la funcionalidad del agente haciendo uso de recursos de la plataforma. Debido a que los bancos podrían requerir que la información que se obtiene de la conversación con un cliente o de un nuevo reclamo generado de manera automatizada no quede alojada en la base de datos de Salesforce y por el contrario se envíe a un sistema externo que forme parte del ecosistema tecnológico del banco, se realizó una prueba a partir de la creación de un servicio de API para que Salesforce comunique los datos obtenidos por el asistente virtual directamente a un sistema externo, simulado con la plataforma de integración Pipedream, la cual permite generar endpoints de prueba para recibir datos de un sistema a través de Internet. En la siguiente imagen se detalla el proyecto de API creado en Pipedream y el flujo de trabajo que permite visualizar un endpoint (punto de entrada de datos) preparado para recibir los datos que el asistente virtual capturó en su conversación con el cliente.

Figura 25

Endpoint Configurado para Recibir Reclamos desde Salesforce



Nota. Obtenido de la configuración de una cuenta de prueba en Pipedream.

El endpoint `https://eo1t4vhhlnqd6x.m.pipedream.net` fue utilizado en el código de Salesforce para permitir informar los detalles de un nuevo reclamo generado. Se crearon dos componentes desarrollados: una clase de Apex que representa el servicio para enviar los datos a otro sistema, y un trigger dentro del objeto Caso (representa un reclamo o consulta que registró el asistente virtual en la base de datos) que permitirá informar los campos seleccionados cada vez que en nuevo caso sea dado de alta. Esta lógica podrá perfeccionarse de acuerdo con lo que requiera cada banco en su conexión con sistemas propios.

Clase Apex principal para ejecutar el servicio web

```
//Clase con Callout para servicio externo
public class CaseCreatedNotifier implements Queueable,
Database.AllowsCallouts {
    private final Set<Id> caseIds;

    // Configuración de Remote Site Settings para el dominio
    public static final String ENDPOINT =
'https://eo1t4vhhlnqd6x.m.pipedream.net';

    public CaseCreatedNotifier(Set<Id> caseIds) {
        this.caseIds = caseIds != null ? caseIds.clone() : new Set<Id>();
    }

    public void execute(QueueableContext qc) {
        if (caseIds.isEmpty()) return;

        List<Case> cases = [
```

```

        SELECT Id, CaseNumber, Subject, Description, CreatedDate,
Origin, ContactId
        FROM Case
        WHERE Id IN :caseIds
    ];

    Http http = new Http();

    for (Case c : cases) {
        // Payload con API names del objeto
        Map<String, Object> payload = new Map<String, Object>{
            'CaseNumber' => c.CaseNumber,
            'Subject'     => c.Subject,
            'Description' => c.Description,
            'CreatedDate' => c.CreatedDate, // ISO 8601 al serializar
            'Origin'      => c.Origin,
            'ContactId'   => c.ContactId
        };

        HttpRequest req = new HttpRequest();
        req.setMethod('POST');
        req.setHeader('Content-Type', 'application/json');
        req.setTimeout(2000); // ms (ajustable)

        req.setEndpoint(ENDPOINT); // Opción A: Remote
Site
        req.setBody(JSON.serialize(payload));

        try {
            HttpResponse res = http.send(req);
            if (res.getStatusCode() >= 400) {
                System.debug(LoggingLevel.WARN,
                    'CaseCreatedNotifier: error ' + res.getStatus() + '
body=' + res.getBody());
            }
        } catch (Exception e) {
            System.debug(LoggingLevel.ERROR, 'CaseCreatedNotifier
exception: ' + e.getMessage());
        }
    }
}

```

La clase CaseCreatedNotifier contiene un proceso asíncrono en Salesforce que, tras la creación de nuevos casos, realiza una llamada HTTP al servicio externo configurado previamente en Pipedream para notificar los datos relevantes del reclamo o consulta. Para ello,

primero una búsqueda recupera el conjunto de casos creados mediante una consulta SOQL y construye un objeto JSON con campos específicos (CaseNumber, Subject, Description, CreatedDate, Origin y ContactId). Luego, envía dicho contenido mediante una solicitud POST al endpoint configurado, utilizando la interfaz Queueable para ejecutar la operación fuera de la transacción principal y Database.AllowsCallouts para habilitar el callout externo. Esta arquitectura asegura la integración con sistemas externos sin afectar la performance ni las restricciones de la operación de inserción del caso.

A su vez, también se incluyó una clase de Apex que representa la clase test con los métodos de prueba diseñados para verificar que la clase principal CaseCreatedNotifier funciona correctamente.

Clase Apex test

```
@IsTest
private class CaseCreatedNotifierTest {

    private class PipedreamMock implements HttpCalloutMock {
        public HTTPResponse respond(HTTPRequest req) {
            System.assertEquals('POST', req.getMethod(), 'Método HTTP
incorrecto');
            System.assertEquals('application/json', req.getHeader('Content-
Type'), 'Header Content-Type');

            // Validación de campos
            Map<String, Object> body =
                (Map<String, Object>)
JSON.deserializeUntyped(req.getBody());

            System.assert(body.containsKey('CaseNumber'), 'Falta
CaseNumber');
            System.assert(body.containsKey('Subject'), 'Falta Subject');
            System.assert(body.containsKey('Description'), 'Falta
Description');
            System.assert(body.containsKey('CreatedDate'), 'Falta
CreatedDate');
            System.assert(body.containsKey('Origin'), 'Falta Origin');
            System.assert(body.containsKey('ContactId'), 'Falta
ContactId');

            HttpResponse res = new HttpResponse();
            res.setStatusCode(200);
            res.setBody('OK');
            return res;
        }
    }
}
```

```

    }
}

@IsTest
static void sendOnInsert() {
    Test.setMock(HttpCalloutMock.class, new PipedreamMock());

    Contact con = new Contact(LastName='Tester',
Email='tester@example.com');
    insert con;

    Case cs = new Case(
        Subject='Prueba de notificación',
        Origin='Email',
        Description='Descripción de prueba',
        ContactId=con.Id
    );

    Test.startTest();
    insert cs; // Ejecuta el trigger -> encola Queueable -> ejecuta
callout en Test.stopTest
    Test.stopTest();
}
}

Trigger en el objeto Caso que permite ejecutar el servicio al sistema externo
trigger CaseAfterInsert on Case (after insert) {
    if (Trigger.isAfter && Trigger.isInsert) {
        System.enqueueJob(new
CaseCreatedNotifier(Trigger.newMap.keySet()));
    }
}

```

Otro desarrollo personalizado del prototipo fue la creación de una acción de control de horario mediante una clase Apex, para que de manera opcional Hori pudiera determinar si el banco se encuentra con atención al público o si se encuentra en una franja horaria fuera de oficina, lo que imposibilita la capacidad de derivar una conversación a un empleado humano en el momento de la consulta del cliente. Esta clase de Apex encapsula la lógica de negocio para comparar la hora actual con el horario de atención y retorna un resultado que el agente utiliza en su diálogo. Es decir que es una funcionalidad de código que brinda soporte al asistente virtual para determinar el horario de atención como información contextual. La integración de Apex se logró gracias a la capacidad nativa de Agentforce de registrar Apex Actions dentro de un agente, lo que permite exponer métodos en código Apex para que sean invocados como

pasos del plan de razonamiento del agente. El motor de razonamiento identifica la necesidad de conocer el horario de atención actual y ejecuta la acción en Apex para obtener los datos. Esta demostración de invocación de lógica en backend presenta la factibilidad técnica de integrar procesos específicos del negocio (en este caso, validación de horarios) en las respuestas del asistente virtual.

Se presenta la utilización de la clase `BusinessHoursIsWithinInvocable` escrita en código Apex, que brinda la posibilidad de validar si una fecha y hora específica se encuentra dentro del horario laboral (Business Hours) configurado en la instancia de trabajo del banco, permitiendo que el asistente virtual conozca el horario de la consulta. La hora de oficina configurada en la instancia de trabajo se corresponde con el horario que determine necesario cada banco. En este caso se determinó que la atención del servicio al cliente para generar reclamos sea desde las ocho de la mañana a seis de la tarde. Es decir que el código, en un flujo que responde a consultas de clientes, permite usar esta acción para decidir si una solicitud llegó dentro del horario de atención o fuera de él. En caso de que se detecte que el horario es fuera de oficina, se podrá indicar al asistente virtual que comunique esto al cliente para que no se esperen respuestas atendidas por un asistente humano con inmediatez, lo que podría generar una falsa expectativa en el cliente.

Clase Apex principal para consulta de horario de atención

```
public with sharing class BusinessHoursIsWithinInvocable {

    // Clase interna para la entrada de la acción invocable
    public class Request {
        @InvocableVariable(required=true,
            label='Business Hours Id',
            description='Id del registro de BusinessHours contra el que se
validará')
        public Id businessHoursId;

        @InvocableVariable(required=false,
            label='Fecha/Hora a evaluar',
            description='Si se omite, se usa System.now()')
        public Datetime whenDT;
    }

    // Clase interna para la salida de la acción invocable
    public class Response {
        @InvocableVariable(label='Dentro de horario',
```

```

        description='True si la fecha/hora dada cae dentro del
BusinessHours')
        public Boolean isWithin;
    }

    /**
     * @InvocableMethod: Verifica si la fecha/hora dada está dentro de los
Business Hours especificados.
     */
    @InvocableMethod(label='¿Dentro de horario? (BusinessHours)',
        description='Devuelve true/false según si la fecha/hora está dentro
del BusinessHours indicado')
    public static List<Response> check(List<Request> requests) {
        List<Response> results = new List<Response>();

        for (Request r : requests) {
            Response res = new Response();

            // 1. Manejo de entradas nulas o incompletas
            if (r == null || r.businessHoursId == null) {
                res.isWithin = false;
                results.add(res);
                // 'Continue' cumple la función de pasar a la siguiente
iteración
                continue;
            }

            // 2. Determinar la fecha/hora a evaluar
            // Si r.whenDT es nulo, usa System.now() (la hora actual del
sistema)
            Datetime ts = (r.whenDT == null) ? System.now() : r.whenDT;

            // 3. Realizar validación con método estándar de Salesforce
            // BusinessHours.isWithin(businessHoursId, targetDateTime)
            res.isWithin = BusinessHours.isWithin(r.businessHoursId, ts);

            results.add(res);
        }

        return results;
    }
}

```

La clase está compuesta por dos estructuras internas: una para la entrada de datos (Request) y otra para la salida (Response). La primera recibe como parámetros el identificador del registro de Business Hours que se desea consultar y una fecha y hora opcional que será

evaluada. En caso de que no se proporcione este valor, el sistema utiliza la hora actual del servidor mediante la función `System.now()`. La segunda estructura define la variable de salida, un valor booleano denominado `isWithin`, que indica si la fecha evaluada se encuentra dentro del horario laboral especificado. El método principal de la clase, denominado `check`, recibe una lista de solicitudes, procesa cada una y devuelve una lista de respuestas. En cada iteración, el método valida que el identificador del horario laboral no sea nulo, determina la fecha a evaluar y utiliza el método estándar `BusinessHours.isWithin()` de Salesforce para verificar la condición. Si la fecha y la hora se encuentran dentro del horario laboral configurado, el resultado devuelto será `true`, de lo contrario, será `false`.

A continuación, se incluye un ejemplo de la clase test para la clase principal de consulta de horario de atención.

Clase test

`@IsTest`

`private class BusinessHoursIsWithinInvocableTest {`

`@IsTest`

`static void testInvocable() {`

`// Business Hours por defecto`

`BusinessHours bh = [`

`SELECT Id FROM BusinessHours WHERE IsDefault = true LIMIT 1`

`];`

`// Request válida`

`BusinessHoursIsWithinInvocable.Request req =`

`new BusinessHoursIsWithinInvocable.Request();`

`req.businessHoursId = bh.Id;`

`// Ejecutar invocable`

`Test.startTest();`

`List<BusinessHoursIsWithinInvocable.Response> res =`

`BusinessHoursIsWithinInvocable.check(new`

`List<BusinessHoursIsWithinInvocable.Request>{ req });`

`Test.stopTest();`

`// validaciones mínimas`

`System.assertEquals(1, res.size());`

`System.assertNotEquals(null, res[0].iswithin);`

`}`

`@IsTest`

`static void testRequestInvalida() {`

```

        // Request sin BusinessHoursId
        BusinessHoursIsWithinInvocable.Request req =
            new BusinessHoursIsWithinInvocable.Request();

        Test.startTest();
        List<BusinessHoursIsWithinInvocable.Response> res =
            BusinessHoursIsWithinInvocable.check(new
List<BusinessHoursIsWithinInvocable.Request>{ req });
        Test.stopTest();

        System.assertEquals(false, res[0].iswithin);
    }
}

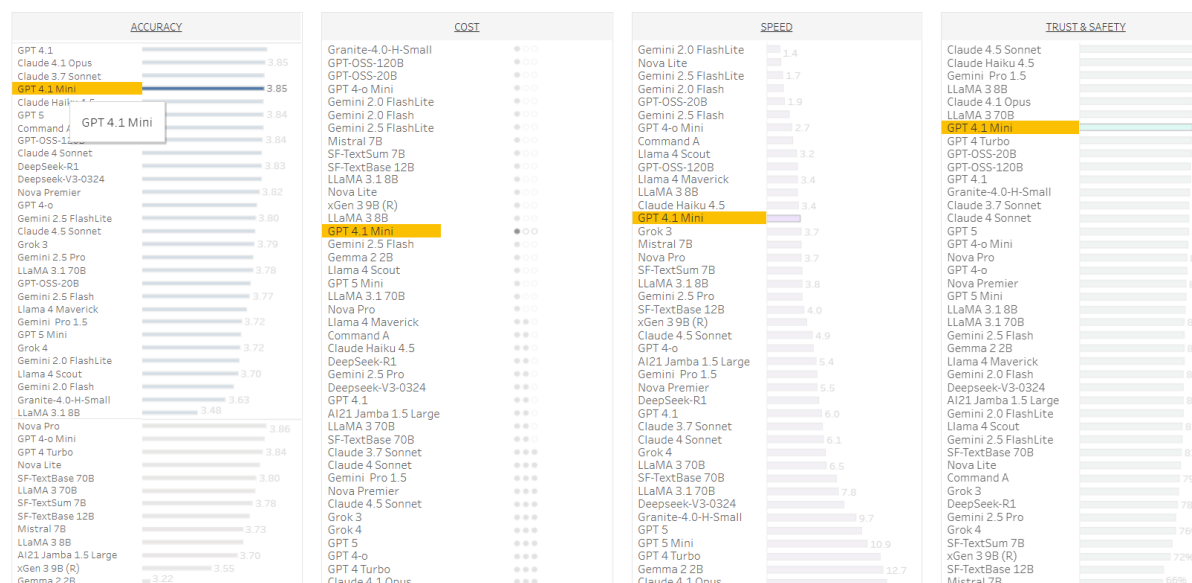
```

Con el objetivo de desplegar una base de conocimiento sobre productos y políticas bancarias internas, se incorporó contenido documental de Banco Horizonte a través de la *Agentforce Data Library* (biblioteca de datos de Agentforce), una funcionalidad que habilita al agente a utilizar datos no estructurados como documentos PDF o artículos redactados para utilizar en sus respuestas. Específicamente, se cargaron cinco artículos con documentos en formato PDF con una guía de atención al cliente, productos y servicios que el banco ofrece a sus clientes, información de seguridad y privacidad, pasos para gestionar trámites y beneficios a los que pueden acceder los clientes.

Luego de crear la información de manera manual en la instancia de trabajo, Salesforce indexó automáticamente esta información para su uso en tiempo real, lo que permite ejecutar una recuperación de datos con apoyo de IA generativa. Esta configuración permitió que Hori pueda responder preguntas consultando directamente el contenido de esos documentos, en lugar de limitarse a un diálogo preprogramado. Por ejemplo, si un cliente pregunta “¿Qué programa de beneficios tiene el Banco?”, el agente busca la respuesta en los artículos cargados sobre beneficios de clientes y proporciona el dato exacto extraído de dicho documento. El asistente virtual con implementación del LLM puede procesar información en texto libre y entender consultas no anticipadas sin entrenamiento específico, apoyándose en el contenido indexado y en las capacidades del LLM para interpretar lenguaje humano, lo cual podrá perfeccionarse a medida que se generen más interacciones. La configuración del prototipo se realizó con la utilización del modelo OpenAI GPT 4.1 Mini, el cual combina rendimiento con bajos costos de consumo de tokens, de acuerdo como se presenta en Salesforce (2025d).

Figura 26

Rendimiento y Costo para GPT 4.1 Mini



Nota. Obtenido de Agentic Benchmark for CRM, Salesforce, 2025d.

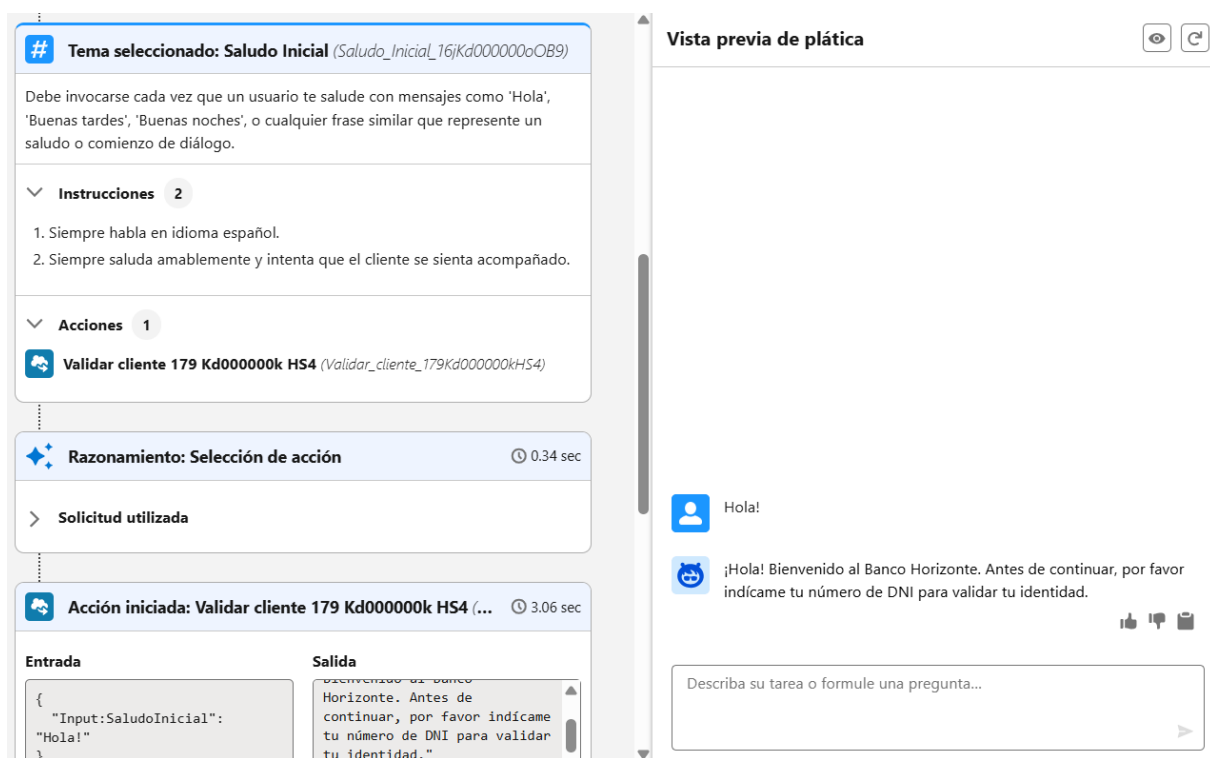
El prototipo permite formular consultas cuyos detalles solo existían en los artículos cargados, obteniendo respuestas correctas basadas en esos documentos. De igual manera se construyó la interacción de Hori con los procesos de atención al cliente. Al residir dentro de la plataforma, el asistente virtual se integra de forma nativa con la operación de soporte, accediendo a los mismos datos de clientes, reclamos y otros datos que los agentes humanos. Agentforce está diseñado para ceder el control a un operador humano cuando enfrenta un requerimiento demasiado complejo o fuera de sus permisos, garantizando que las consultas sean resueltas eficientemente con intervención más especializada. En la configuración de Hori se establecieron criterios para dicha derivación. Cuando se reciben ciertas palabras clave o intentos fallidos repetidos, se activa una sugerencia de crear un reclamo en caso de que el cliente lo confirme. Gracias a la infraestructura unificada, toda la información recopilada por el asistente virtual durante la interacción (contexto del cliente, intención detectada, acciones realizadas) puede traspasarse a la base de datos. Hori puede crear automáticamente una instancia del objeto Caso de Salesforce con el resumen de la conversación si no logró resolver un problema presentado por un cliente, de modo que un agente humano luego trabaje con todos los antecedentes necesarios. Si bien en el prototipo no se implementó una consola de supervisor completa, la capacidad inherente de activar logs en cada acción detectada y ejecutada por el asistente virtual refuerza la viabilidad de adoptar a Hori en un entorno productivo.

5.4 Resultados de las Pruebas

A medida que se fueron configurando los prompts, tópicos y acciones, se utilizó el entorno de prueba de Agentforce para simular conversaciones y validar su comportamiento. La plataforma incluye un simulador de conversaciones interactivo que permitió enviar distintos mensajes de ejemplo a Hori y observar cómo el agente planificaba su respuesta paso a paso, de la misma manera que se presenta en la siguiente imagen en la que se inicia una conversación con un usuario de tipo cliente y se le solicita su validación de identidad para continuar.

Figura 27

Vista Previa de Respuesta Inicial al Cliente



Nota. Elaboración propia en el simulador de conversaciones de Salesforce.

Sobre la parte derecha de la imagen se visualiza la conversación mantenida entre el asistente virtual y un posible cliente, quien ejecutó de manera inicial el mensaje “Hola”. Sobre la izquierda de la imagen se presentan los pasos que el asistente virtual ejecutó para resolver una respuesta con el saludo inicial que da comienzo a la conversación.

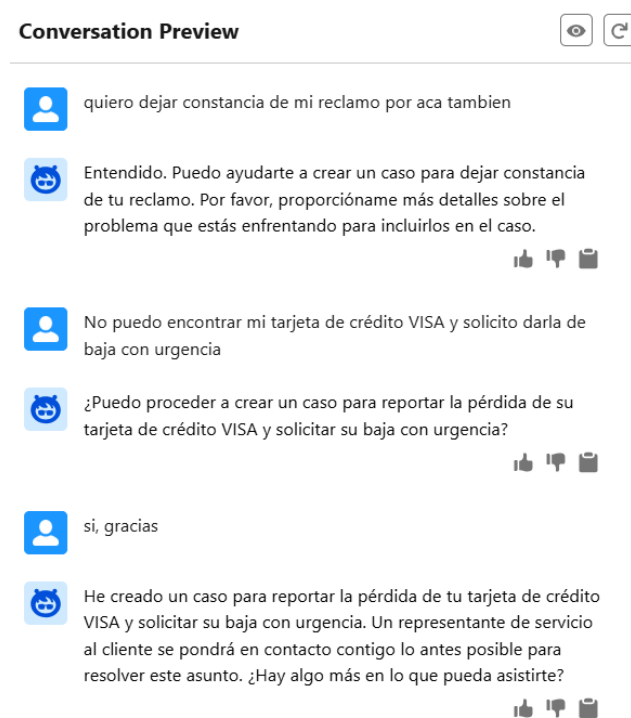
A diferencia de chatbots tradicionales, no fue necesario programar frases de entrenamiento ni sinónimos para cada intención: el motor de razonamiento con lenguaje natural que interactúa con el modelo de GPT, comprende las consultas de forma autónoma gracias al modelo generativo. De esta manera, Hori puede interpretar una pregunta formulada de

múltiples maneras. Por ejemplo, el asistente virtual puede comprender qué debe responder con los horarios de atención en las sucursales del banco si se preguntara: “¿Está abierto el banco ahora?”, o de igual manera si un usuario preguntara: “¿A qué hora cierran las sucursales?”, sin definir manualmente cada variación, ya que comprende la intención subyacente y accede a la información relevante.

Con respecto a la generación de reclamos, el asistente virtual cuenta con un tópico configurado para generar reclamos en la base de datos interna a partir de preguntar al usuario cliente una descripción del problema, como se presenta en la siguiente conversación. La imagen presenta la conversación observada desde el componente de previsualización que Salesforce ofrece para probar el funcionamiento de los agentes creados.

Figura 28

Registro de Reclamo Automatizado



Nota. Elaboración propia en pruebas con el asistente virtual.

La acción automatizada del asistente virtual generó un nuevo registro de reclamo en la base de datos de Salesforce, lo que permitirá que luego un asistente humano del banco pueda gestionar el problema del cliente por fuera de la conversación original. En la siguiente imagen se presenta el registro del pedido de asistencia del cliente.

Figura 29

Visualización del Reclamo en la Base de Datos

The screenshot shows a Salesforce case record. At the top, there's a header bar with a red square icon containing a white document symbol, followed by the text 'Caso' and 'Pérdida de tarjeta VISA'. Below this, there's a section with three fields: 'Prioridad' (Medium), 'Estado' (New, with a yellow star icon), and 'Número del caso' (00001385). The main body of the record is titled 'Detalles' and contains several fields arranged in two columns. The left column includes 'Propietario del caso' (General), 'Asunto' (Pérdida de tarjeta VISA), 'Descripción' (El cliente reporta la pérdida de su tarjeta de crédito VISA y solicita darla de baja con urgencia para evitar cualquier uso no autorizado.), and 'Fecha/Hora de apertura' (11/9/2025, 9:04 p. m.). The right column includes 'Número del caso' (00001385), 'Nombre del contacto' (Carlos Lopez, with a link icon), 'Origen del caso' (Web), and 'Email Web' (carlos.lopez@ejemplo.com, with a link icon). There are also small edit icons (pencil) next to the 'Asunto' and 'Descripción' fields.

Caso	
Pérdida de tarjeta VISA	
Prioridad	Estado
Medium	★ New
Número del caso	
00001385	
Detalles	
Propietario del caso	Número del caso
General	00001385
▼ Detalles	
Asunto	Nombre del contacto
Pérdida de tarjeta VISA	Carlos Lopez
Descripción	Origen del caso
El cliente reporta la pérdida de su tarjeta de crédito VISA y solicita darla de baja con urgencia para evitar cualquier uso no autorizado.	Web
Fecha/Hora de apertura	Email Web
11/9/2025, 9:04 p. m.	carlos.lopez@ejemplo.com

Nota. Elaboración propia a partir de la interfaz de visualización de la base de datos.

Así como la información proporcionada por el cliente se puede guardar directamente en la plataforma, también puede ser comunicada a un sistema externo del banco para continuar con la ejecución de un flujo más extenso. Para ejecutar la acción desarrollada que permite enviar un mensaje saliente por integración a un sistema externo desde Salesforce, se debe recibir la solicitud de un cliente de crear un nuevo reclamo en su conversación con el asistente virtual, tal como se presenta en la siguiente imagen. Como el cliente fue validado previamente, se conocerán sus datos de contacto y se podrá registrar un caso en Salesforce con el reclamo pertinente.

Figura 30

Conversación para Solicitar un Reclamo

 quiero cargar un reclamo porque tengo un gasto no autorizado en mi tarjeta de credito

 ¿Puedo proceder a generar un reclamo por el gasto no autorizado en tu tarjeta de crédito?

 si

 He generado un reclamo por el gasto no autorizado en tu tarjeta de crédito. ¿Hay algo más en lo que pueda asistirte?

Nota. Elaboración propia en pruebas con el asistente virtual.

En la conversación se presenta una mejora de respuesta con respecto al caso anterior, dado que el asistente virtual utiliza la palabra reclamo para acercar su lenguaje de una manera más natural al cliente. La información proporcionada por el cliente (representado por el campo Nombre del contacto) son datos que fueron incluidos en la creación de un reclamo con el número 1386 guardado directamente de manera automática en la base de datos de Salesforce.

Figura 31

Nuevo Reclamo generado con Asistente Virtual


Caso
Gasto no autorizado en tarjeta

Prioridad	Estado	Número del caso
Medium	★ New	00001386

Detalles

Propietario del caso	Número del caso
General	00001386

▼ Detalles

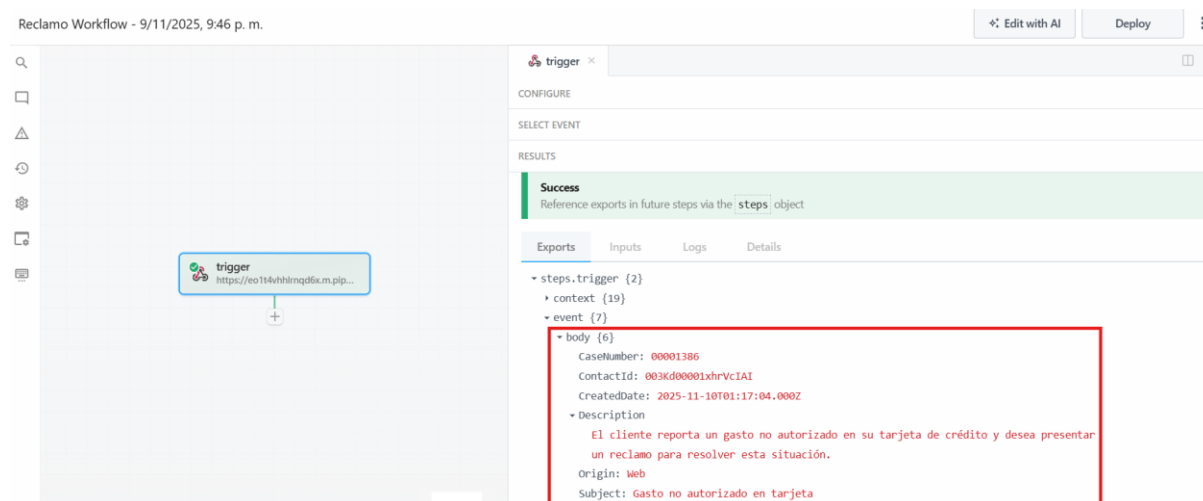
Asunto	Nombre del contacto
Gasto no autorizado en tarjeta	Carlos Lopez
Descripción	Origen del caso
El cliente reporta un gasto no autorizado en su tarjeta de crédito y desea presentar un reclamo para resolver esta situación.	Web
Fecha/Hora de apertura	Email Web
11/9/2025, 10:17 p. m.	carlos.lopez@ejemplo.com

Nota. Obtenido de la instancia propia de trabajo en Salesforce.

El funcionamiento de la comunicación con un sistema externo mediante la ejecución del código Apex desarrollado se presenta en la siguiente imagen con la recepción del mensaje en Pipedream, el cual contiene directamente la información recuperada por el asistente virtual en su conversación con el cliente y que incluye los datos más importantes del reclamo número 1386 que fue creado anteriormente.

Figura 32

Recepción del Mensaje en Sistema Externo Simulado



Nota. Obtenido del mensaje POST recibido desde Salesforce hacia Pipedream.

5.5 Conclusiones del Prototipo

El prototipo presentado permitió no solo construir y probar a Hori de forma aislada, sino también verificar que el asistente virtual utiliza los objetos y flujos activos en Salesforce como la información de cliente, la validación de cliente, reclamos y los artículos de procedimientos internos del banco, y puede coexistir con integraciones con otros sistemas de un banco comercial, lo que facilita su futura evolución hacia un piloto en producción con mínimo esfuerzo de integración.

Desde el punto de vista técnico, la Developer Edition ofreció un entorno controlado y suficiente para experimentar con Agentforce, Service Cloud y Data Cloud, a pesar de sus límites de funcionalidades y almacenamiento. Las pruebas realizadas confirman que esta instancia de trabajo replica las capacidades centrales de una organización empresarial, habilitando objetos como entidades de una base de datos, API y desarrollo en código Apex que permite integrar al asistente virtual con procesos de negocio básicos y extender su funcionalidad hacia otros sistemas externos.

Este diseño funcional prototipado será transferible a un piloto productivo siempre que se acompañe su implementación con un proyecto con mecanismos de gobernanza y control de consumo propios de soluciones con modelos de lenguaje. En particular, la supervisión de tokens y costos asociados debe formar parte de la operación desde el inicio, con alertas y cuotas razonables para evitar desvíos por picos de demanda o interacciones anómalas. Este aspecto fue considerado en el capítulo 4 en la propuesta como requisito operativo para lograr una adopción progresiva. Asimismo, la experiencia del prototipo permitió identificar lineamientos de diseño seguro que deberán mantenerse en un despliegue real: acceso a una plataforma segura, uso de objetos nativos (Casos, Contactos, Artículos) y desarrollo de código siguiendo las buenas prácticas de la programación orientada a objetos.

En cuanto a las limitaciones, el alcance del prototipo permitió ejecutar pruebas de consultas frecuentes y apertura de reclamos, sin exponer hacia el cliente datos transaccionales reales. También se trabajó con volúmenes acotados de datos, artículos de procedimientos de banco simulados y reglas simples de validación, suficientes a instancias de demostración que deberán ser profundizados en una implementación completa del asistente virtual para atención al cliente en bancos argentinos, en línea con la propuesta presentada en el capítulo 4.

Conclusiones

Los chatbots actuales utilizados en la atención al cliente en el sector bancario argentino presentan limitaciones por su desarrollo basado en reglas predefinidas. En por ello que se propuso la implementación de un asistente virtual inteligente con procesamiento de lenguaje natural con la aplicación de un modelo de lenguaje de gran escala a través de la plataforma Salesforce, que permita satisfacer la demanda de inmediatez y personalización que exigen los clientes.

Para consolidar una propuesta innovadora, el trabajo consideró experiencias previas de automatización conversacional en bancos nacionales e internacionales, y el impacto de la incorporación de la plataforma Salesforce con componentes como Data Cloud, Service Cloud y Agentforce, y mecanismos de recuperación aumentada (RAG). El asistente virtual propuesto presenta los principales flujos de negocio de la banca minorista: consultas básicas, gestión de reclamos, gestiones de seguridad digital y revisión de transacciones financieras. El uso de RAG para ejecutar estos procesos en línea con clientes, constituye un valor diferencial con respecto a los chatbots tradicionales, al permitir que el asistente virtual fundamente sus respuestas en datos validados por el banco y genere respuestas novedosas a medida de cada conversación.

El prototipo funcional desarrollado en la plataforma Salesforce presenta la configuración de prompts iniciales, validación de identidad del cliente, acciones declarativas y clases en código Apex, componentes que demuestran la capacidad de la plataforma para ejecutar un flujo conversacional completo dentro del ecosistema bancario, haciendo uso de la aplicación de modelos de lenguaje a gran escala. Además, la integración con un endpoint externo y la consulta de artículos de una base de conocimiento interna consolidaron un circuito de respuesta automatizada para prácticas de atención bancaria reales, en vinculación con sistemas bancarios. El asistente propuesto no reemplaza el criterio profesional humano, sino que lo complementa, permitiendo elevar la calidad en la atención al cliente minorista y habilitando a que las personas se concentren en tareas de mayor complejidad. Esto permite cumplir con objetivos de negocio como la reducción de los tiempos de espera en la atención, el aumento en la precisión de las respuestas a clientes y la posibilidad de ofrecer un soporte multicanal unificado.

La adopción de la solución por parte de los bancos implica desafíos que exceden las cuestiones técnicas. La transición hacia esquemas basados en inteligencia artificial generativa requiere procesos de capacitación, prácticas de gobierno de datos robustas y un compromiso con la transparencia en la interacción con los usuarios. La experiencia analizada muestra que

estos factores condicionan la viabilidad de la propuesta y constituyen líneas concretas para futuros desarrollos, especialmente en relación con la integración de nuevas fuentes de datos y la evaluación continua del desempeño del asistente virtual. En la misma línea, la viabilidad de la propuesta exige una adecuada adopción por parte de las personas, tanto de aquellas que forman parte de la estructura interna de los bancos como de los clientes que utilizarán el asistente virtual como punto de contacto. Los líderes tecnológicos de los bancos son actores fundamentales, ya que deben trasladar a los equipos de trabajo cómo incorporar estas herramientas a su práctica diaria y reconocerlas como un apoyo que facilita su trabajo. La automatización, en este sentido, no puede concebirse de manera aislada, sino como parte de un proceso más amplio de transformación digital que involucra una cultura organizacional con capacitación y confianza en la tecnología.

Esta propuesta no solo responde a las necesidades actuales de la banca argentina, sino que también abre la posibilidad de ser adaptada a otros sectores donde la atención al cliente requiere altos niveles de automatización y precisión. Con ello, se busca sentar las bases para una adopción progresiva de tecnologías de inteligencia artificial generativa que complementen el rol humano en la gestión de servicios profesionales.

Trabajos Futuros

Se identifican diversas líneas de desarrollo que podrían ampliar la propuesta de implementación del asistente virtual. Una primera línea de trabajo consiste en avanzar hacia un modelo de interacción conversacional multimodal, que no se limite al procesamiento de texto, sino que incorpore voz, imagen y video como canales complementarios de comunicación. Esta evolución permitiría que los clientes interactúen con el asistente mediante comandos de voz o compartiendo archivos que brinden una contextualización más precisa.

En segundo lugar, se propone investigar mecanismos de aprendizaje continuo y adaptación dinámica del modelo, mediante la retroalimentación de las interacciones reales con los usuarios. La integración de un ciclo de mejora basado en métricas de desempeño (precisión, satisfacción del cliente, tasa de resolución) permitiría que el asistente refine sus respuestas y sugerencias de manera autónoma, bajo supervisión humana.

Otra línea relevante de investigación futura se vincula con la evaluación empírica de impacto organizacional. Es decir, estudiar cómo la adopción de asistentes virtuales inteligentes incide en los indicadores de productividad, en la redistribución de tareas del personal de atención y en la percepción de los clientes respecto de la calidad del servicio brindado. Estos estudios permitirían obtener evidencia cuantitativa sobre los beneficios operativos de la automatización conversacional y contribuirían a definir marcos de adopción gradual en instituciones financieras.

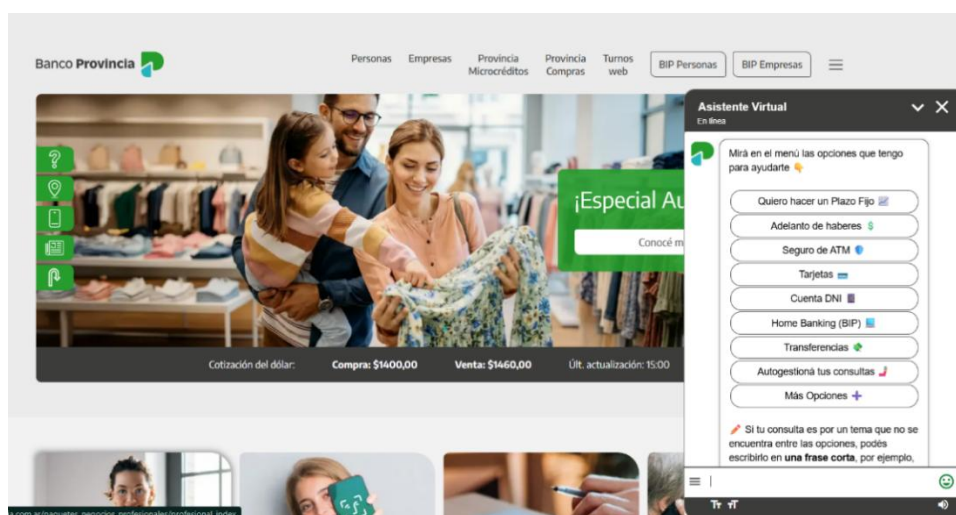
Se propone también continuar investigando la dimensión ética y regulatoria de la inteligencia artificial generativa en el sector financiero. En especial, se plantea profundizar en la aplicación de marcos de IA responsable y auditorías de sesgo en los modelos de lenguaje utilizados. A medida que la tecnología evolucione y las normativas locales e internacionales se actualicen, será fundamental que las entidades bancarias cuenten con políticas claras sobre uso de datos, supervisión humana y límites de automatización, lo cual deberá llevar más tiempo de análisis que el contenido en el cronograma de esta propuesta.

Anexo

Para analizar las implementaciones actuales de asistentes virtuales en bancos de argentina se experimentaron conversaciones con los chatbots activos. A continuación, se presentan las imágenes de las interacciones con los asistentes virtuales de cada banco.

Figura 33

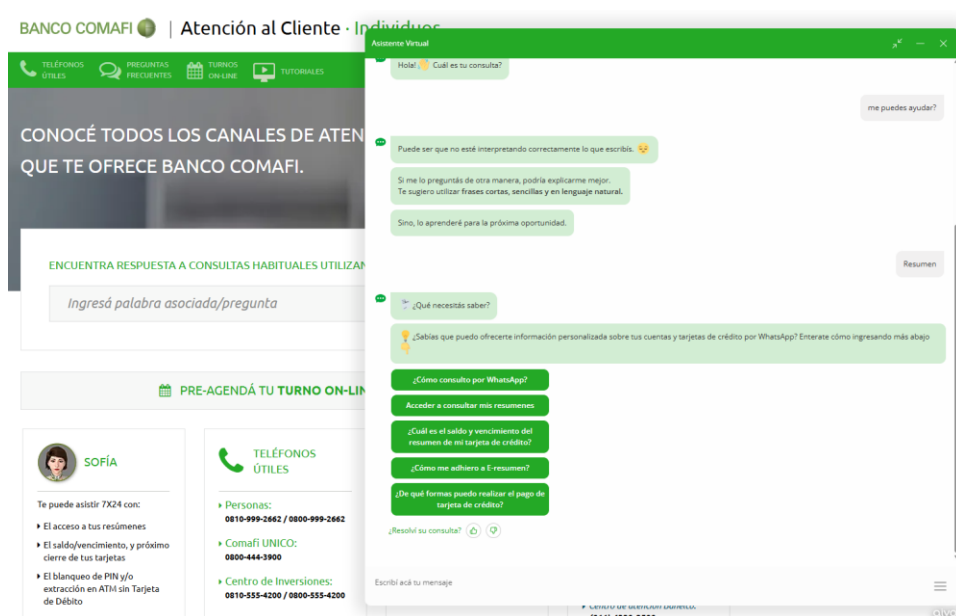
Conversación con chat online del Banco Provincia



Nota. Elaboración propia.

Figura 34

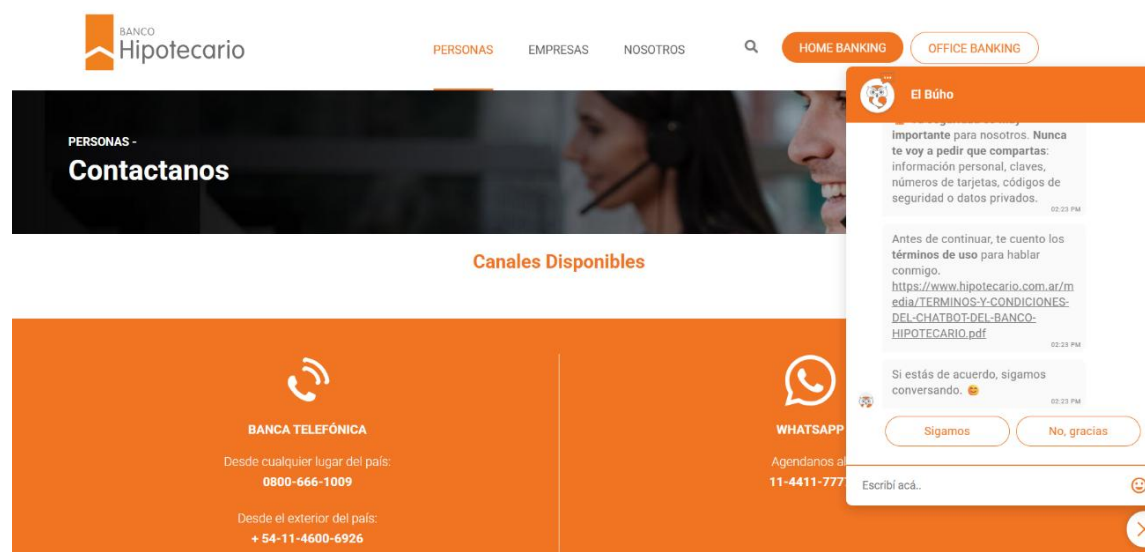
Conversación de chat del Banco Comafi



Nota. Elaboración propia.

Figura 35

Conversación con El Búho, chat online del Banco Hipotecario



Nota. Elaboración propia

Figura 36

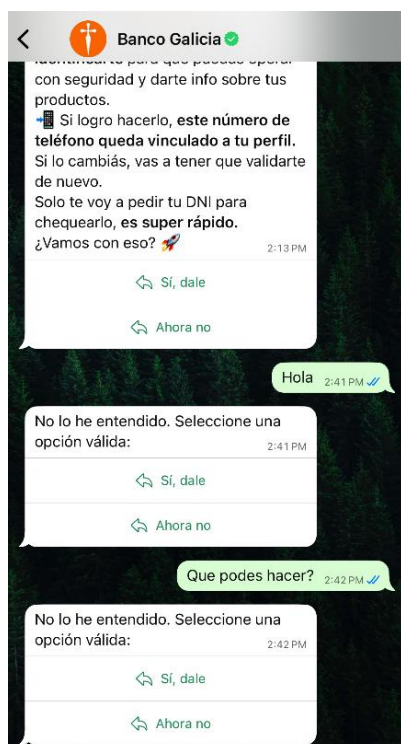
Conversación con chatbot del Banco Ciudad



Nota. Elaboración propia.

Figura 37

Conversación con Banco Galicia

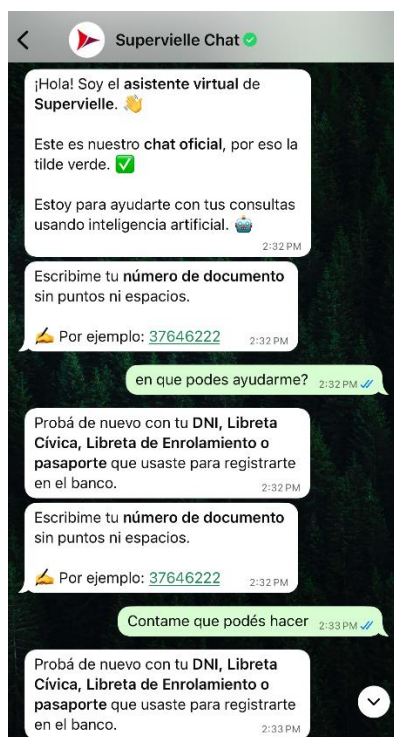
*Nota.* Elaboración propia.**Figura 38**

Conversación con BancoChat del Banco Macro

*Nota.* Elaboración propia.

Figura 39

Conversación con chat del Banco Supervielle



Nota. Elaboración propia.

Acrónimos

IA	Inteligencia Artificial
API	Application Programming Interface
BPMN	Business Process Model and Notation
CRM	Customer Relationship Management
KPI	Key Performance Indicator
LLM	Large Language Model
NPS	Net Promoter Score
PLN	Procesamiento de Lenguaje Natural
RAG	Retrieval-Augmented Generation
SLA	Service Level Agreement
UX	User Experience

Referencias

- Acceso a Drive público con la bibliografía utilizada:
https://drive.google.com/drive/folders/16J_-Hj5puJ0UCWLpoDjybP6cyyQdVsCw?usp=sharing
- Accenture. (2025). Banking consumer study 2025: Where is the love? How advocacy drives loyalty and organic growth in banking. [Estudio del consumidor bancario 2025: ¿Dónde está el amor? Cómo la promoción impulsa la fidelización y el crecimiento orgánico en la banca.] <https://www.accenture.com/lk-en/insights/banking/consumer-study-banking-advocacy-powering-growth>
- Aivo. (2025). Caso de Éxito Banco Comafi. <https://es.aivo.co/customer-stories/banco-comafi>
- Alammar, J., & Grootendorst, M. (2024). Hands-on large language models: Language understanding and generation. [Práctica con modelos lingüísticos de gran escala: comprensión y generación del lenguaje.]
- Ámbito. (2025). De los chatbots a la IA generativa: así se preparan los bancos argentinos para el futuro. <https://www.ambito.com/finanzas/de-los-chatbots-la-ia-generativa-asi-se-preparan-los-bancos-argentinos-el-futuro-n6180556>
- Apiux. (2025). Productos digitales automáticos y creación de IA generative. <https://apiux.com/caso-de-exito/productos-digitales-automaticos-y-creacion-de-ia-generativa/>
- Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. [Un modelo probabilístico del lenguaje basado en redes neuronales.] <https://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf>
- Bigeard, A., Nashold, L., Krishnan, R., & Wu, S. (2025). Finance Agent Benchmark: Benchmarking LLMs on real-world financial research tasks [Finance Agent Benchmark: Evaluación de LLM en tareas reales de investigación financiera]. <https://arxiv.org/abs/2508.00828>
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... y Liang, P. (2021). On the opportunities and risks of foundation models. Stanford Center for Research on Foundation Models. [Sobre las oportunidades y los riesgos de los

-
- modelos de cimentación. Centro de Investigación sobre Modelos de Cimentación de Stanford.] <https://crfm.stanford.edu/assets/report.pdf>
- Burkov, A. (2023). The hundred-page language models book. [El libro de modelos lingüísticos de cien páginas.]
- Business Insider. (2025). Morgan Stanley and Bank of America are focusing AI power on tools to make employees more efficient. [Morgan Stanley y Bank of America están concentrando el poder de la IA en herramientas para hacer que los empleados sean más eficientes.] <https://www.businessinsider.com/internal-artificial-intelligence-solutions-banking-companies-employee-training-2025-5>
- Cath, C. (2018). Governing artificial intelligence: Ethical, legal and technical opportunities and challenges [Gobernar la inteligencia artificial: Oportunidades y desafíos éticos, legales y técnicos] <http://doi.org/10.1098/rsta.2018.0080>
- Chen, B., Wu, Z., & Zhao, R. (2023). From fiction to fact: the growing role of generative AI in business and finance. *Journal of Chinese Economic and Business Studies*. [De la ficción a la realidad: el creciente papel de la IA generativa en los negocios y las finanzas. *Revista de Estudios Económicos y Empresariales de China*.] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4528225
- Dignum, V. (2019). Responsible artificial intelligence: How to develop and use AI in a responsible way [Inteligencia artificial responsable: Cómo desarrollar y usar la IA de manera responsable] <https://link.springer.com/book/10.1007/978-3-030-30371-6>
- Eisenstein, J. (2019). Introduction to Natural Language Processing. [Introducción al procesamiento del lenguaje natural].
- Elman, J. L. (1990). Finding structure in time. [Encontrando estructura en el tiempo.] https://doi.org/10.1207/s15516709cog1402_1
- Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). RAGAS: Automated Evaluation of Retrieval Augmented Generation. [RAGAS: evaluación automatizada de la generación aumentada con recuperación.] <https://aclanthology.org/2024.eacl-demo.16.pdf>

-
- Fernández, A. (2019). Inteligencia artificial en los servicios financieros. <https://repositorio.bde.es/handle/123456789/8448>
- Finardi, P., Viegas, J. D., Ferreira, G. T., Mansano, A. F., & Caridá, V. F. (2021). BERTaú: Itaú BERT for digital customer service. [BERTaú: Itaú BERT para atención al cliente digital.] <https://arxiv.org/abs/2101.12015>
- Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society [Un marco unificado de cinco principios para la IA en la sociedad] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3831321
- Galileo Financial Technologies. (2025a). Guía de la IA en la banca latinoamericana: 2025. <https://www.galileo-ft.com/es/blog/ia-en-la-banca-america-latina-2025-guia/>
- Galileo Financial Technologies. (2025b). La revolución digital del sector bancario argentino. <https://www.galileo-ft.com/es/blog/revolucion-digital-sector-bancario-argentino/>
- Galileo Financial Technologies y PYMNTS. (2025). Embedded finance tracker: Beyond the bot. Why embedded conversational AI is banking's next strategic advantage. [Rastreador financiero integrado: Más allá del bot. Por qué la IA conversacional integrada es la próxima ventaja estratégica de la banca.] https://www.pymnts.com/?post_type=tracker_posts&p=2751514
- García Molina, J., Ortín, M. J., Moros, B., Nicolás, J., & Toval, A. (2006). De los procesos del negocio a los casos de uso (UML)
- Golani, R. R. (2025). LLM fine-tuning vs prompt engineering for consumer products. International Journal on Science and Technology. [Ajuste fino de LLM frente a ingeniería de prompts para productos de consumo.] <https://www.ijstat.org/papers/2025/2/3098.pdf>
- Goldberg, Y. (2017). Neural network methods for natural language processing. Morgan & Claypool. [Métodos de redes neuronales para el procesamiento de lenguaje natural].
- Gotts, P., King, R., & Lavous, M. (2024). Elements: The ultimate guide to creating agents [Elementos: La guía definitiva para crear agentes]

-
- Guan, Y., Wang, D., Chu, Z., Wang, S., Ni, F., Song, R., Li, L., Gu, J., & Zhuang, C. (2023). Intelligent virtual assistants with LLM-based process automation. [Asistentes virtuales inteligentes con automatización de procesos basada en LLM.] <https://arxiv.org/abs/2312.06677>
- Gökdemir, O., Siebenschuh, C., Brace, A., Wells, A., Hsu, B., Hippe, K., Setty, P. V., Ajith, A., Pauloski, J. G., Sastry, V., Foreman, S., Zheng, H., Ma, H., Kale, B., Chia, N., Gibbs, T., Papka, M. E., Brettin, T., Alexander, F. J., ... Ramanathan, A. (2025). HiPerRAG: High-performance retrieval augmented generation for scientific insights. [HiPerRAG: Generación aumentada con recuperación de alto rendimiento para descubrimientos científicos.] <https://arxiv.org/abs/2505.04846>
- Hernández Sampieri, R. (2014). Metodología de la Investigación.
- Hochreiter, S., Schmidhuber, J. (1997). Long Short-Term Memory. Neural Comput. [Memoria a largo plazo. Computación neuronal.] <https://www.bioinf.jku.at/publications/older/2604.pdf>
- IBM. (s.f.a). Datos estructurados vs datos no estructurados: ¿cuál es la diferencia? <https://www.ibm.com/es-es/think/topics/structured-vs-unstructured-data>
- IBM. (s. f.b). ¿Qué es el prompt engineering? <https://www.ibm.com/es-es/think/topics/prompt-engineering>
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines [El panorama global de las directrices éticas de la IA] <https://arxiv.org/pdf/1906.11668>
- Jurafsky, D., & Martin, J. H. (2023). Speech and language processing. [Procesamiento del habla y del lenguaje]. https://web.stanford.edu/~jurafsky/slp3/ed3book_aug25.pdf
- Kihn, M. (2025). Agentforce: Harnessing the agency of AI to scale, grow, and lead any industry [Agentforce: Aprovechando la agencia de la IA para escalar, crecer y liderar cualquier industria]
- Koehn, P. (2010). Statistical machine translation. [Traducción automática estadística].

-
- Lenovo. (2025). ThinkPad E14 Gen 7 (14" Intel) - Especificaciones y precios en Argentina. <https://www.lenovo.com/ar/es/p/notebooks/thinkpad/thinkpad-e/lenovo-thinkpad-e14-gen-7-14-inch-intel/21t9cto1wwar2>
- Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Acosta-Navas, D., Hudson, D. A., ... Koreeda, Y. (2023). Holistic Evaluation of Language Models. *Transactions on Machine Learning Research*. [Evaluación Holística de Modelos Lingüísticos. Transacciones sobre investigación sobre aprendizaje automático.]
- Lu, G., Guo, X., Zhang, R., Zhu, W., & Liu, J. (2025). BizFinBench: A business-driven real-world financial benchmark for evaluating LLMs [BizFinBench: Un conjunto de evaluación financiero del mundo real, impulsado por negocios, para medir LLM]. <https://arxiv.org/abs/2505.19457>
- Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing. [Fundamentos del procesamiento de lenguaje natural estadístico]. https://icog-labs.com/wp-content/uploads/2014/07/Christopher_D._Manning_Hinrich_Schütze_Foundations_Of_Statistical_Natural_Language_Processing.pdf
- Martel, A. (2014). Gestión práctica de proyectos con Scrum: Desarrollo de software y gestión ágil de proyectos.
- Mashaabi, A., Almutadha, Y., Abualkishik, A. Z., & Gani, A. (2022). Natural language processing in customer service: A systematic review [Procesamiento de lenguaje natural en el servicio al cliente: Una revisión sistemática] <https://arxiv.org/abs/2212.09523>
- McCarthy, J. (2004). What is artificial intelligence? [¿Qué es la inteligencia artificial?] <https://cse.unl.edu/~choueiry/S09-476-876/Documents/whatisai.pdf>
- McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1956). A proposal for the Dartmouth summer research project on artificial intelligence. [Una propuesta para el

-
- proyecto de investigación de verano de Dartmouth sobre inteligencia artificial.] <http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf>
- McTear, M. F. (2020). Conversational AI: Dialogue systems, conversational agents, and chatbots. [IA conversacional: sistemas de diálogo, agentes conversacionales y chatbots].
- Meduri, S. (2024). Revolutionizing customer service: The impact of large language models on chatbot performance [Revolucionando la atención al cliente: El impacto de los modelos de lenguaje en el desempeño de chatbots] <https://ijsrceit.com/index.php/home/article/view/CSEIT241051057>
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. <https://arxiv.org/abs/1301.3781v3>
- Oche, A. J., Folashade, A. G., Ghosal, T., & Biswas, A. (2025). A systematic review of key retrieval-augmented generation (RAG) systems: Progress, gaps, and future directions. [Una revisión sistemática de los principales sistemas de generación aumentada con recuperación (RAG): avances, vacíos y direcciones futuras.] <https://arxiv.org/abs/2507.18910>
- OpenText (2024). Digital Banking Report: State of AI in Banking [Reporte digital de banca: Estado de la IA en la banca]. <https://www.opentext.com/media/report/state-of-ai-in-banking-digital-banking-report-en.pdf>
- Project Management Institute. (2013). Guía de los fundamentos para la dirección de proyectos.
- Russell, S., & Norvig, P. (2021). Artificial intelligence: A modern approach. [Inteligencia artificial: un enfoque moderno.]
- Salesforce (s.f.a). What is Agentforce? [¿Qué es Agentforce?] <https://www.salesforce.com/au/agentforce/>
- Salesforce. (s.f.b). Einstein Trust Layer: Designed for Trust. [Capa de confianza de Einstein: Diseñada para la confianza.] https://help.salesforce.com/s/articleView?id=ai.generative_ai_trust_arch.htm&type=5

Salesforce. (s. f.c). Select Identity Resolution Match Rules [Seleccionar reglas de coincidencia para la resolución de identidades]. <https://trailhead.salesforce.com/content/learn/modules/data-and-identity-in-salesforce-cdp/select-identity-resolution-match-rules>

Salesforce. (s. f.d). Get Started with Identity Resolution Rulesets [Comenzar con los conjuntos de reglas de resolución de identidades]. <https://trailhead.salesforce.com/content/learn/projects/quick-start-create-identity-resolution-ruleset/get-started-with-identity-resolution-ruleset>

Salesforce. (s.f.e). When to Use Data Loader. [Cuándo utilizar Data Loader.] https://developer.salesforce.com/docs/atlas.en-us.dataLoader.meta/dataLoader/when_to_use_the_data_loader.htm

Salesforce. (2024a). Apex Developer Guide. [Guía para desarrolladores de Apex.] https://developer.salesforce.com/docs/atlas.en-us.apexcode.meta/apexcode/apex_dev_guide.htm

Salesforce. (2024b). Introducing Agentforce 2.0: The Digital Labor Platform for Building a Limitless Workforce. [Presentamos Agentforce 2.0: la plataforma laboral digital para construir una fuerza laboral ilimitada.] <https://www.salesforce.com/news/stories/agentforce-boosts-customer-experience-sales/salesforce-customers-turn-ai-hype-into-reality-with-agentforce>

Salesforce. (2024c). Salesforce named a Leader in the 2024 Gartner® Magic Quadrant™ for Sales Force Automation Platforms. [Salesforce nombrado líder en el Cuadrante Mágico™ de Gartner® 2024 para plataformas de automatización de la fuerza de ventas.] <https://www.salesforce.com/news/stories/gartner-magic-quadrant-sales-force-automation-2024/>

Salesforce. (2024d). Salesforce unveils Agentforce – what AI was meant to be [Salesforce presenta Agentforce - Lo que la IA debía ser] <https://www.salesforce.com/news/press-releases/2024/09/12/agentforce-announcement>

Salesforce. (2024e). Salesforce’s Agentforce is here: Trusted, autonomous AI agents to scale your workforce. [Agentforce de Salesforce ya está aquí: agentes de IA autónomos y

confiables para escalar su fuerza laboral.] <https://www.salesforce.com/news/press-releases/2024/10/29/agentforce-general-availability-announcement>

Salesforce. (2025a). Introducing the Agentforce guide. [Introducción a la Guía de Agentforce.] <https://www.salesforce.com/agentforce/guide/>

Salesforce. (2025b). Agentforce: Planificación de agentes. Realizar prototipos del agente. <https://trailhead.salesforce.com/es/content/learn/modules/agentforce-agent-planning/prototype-the-agent>

Salesforce. (2025c). Agentforce pricing. [Lista de Precios de Agentforce] <https://www.salesforce.com/agentforce/pricing/>

Salesforce. (2025d). Agentic Benchmark for CRM. [Referencia de agentes para CRM.] <https://www.salesforceairesearch.com/crm-benchmark>

Salesforce. (2025e). Asistente de servicios bancarios: caso de uso de Agentforce. <https://www.salesforce.com/mx/agentforce/use-cases/financial-services/banking-services-assistant/>

Salesforce. (2025f). Data Cloud Consumption Estimate. [Estimación del consumo de datos en la nube.]

Salesforce. (2025g). Financial Services Cloud User Guide. [Guía del usuario de Financial Services Cloud.] https://resources.docs.salesforce.com/latest/latest/en-us/sfdc/pdf/financial_services.pdf

Salesforce. (2025h). How does Data Cloud work? [¿Cómo funciona Data Cloud?] <https://www.salesforce.com/data/how-it-works/>

Salesforce. (2025i). Salesforce launches Agentforce 2dx with new capabilities to embed proactive agentic AI into any workflow. [Salesforce lanza Agentforce 2dx con nuevas capacidades para integrar IA proactiva en cualquier flujo de trabajo.] <https://investor.salesforce.com/news/news-details/2025/Salesforce-Launches-Agentforce-2dx-with-New-Capabilities-to-Embed-Proactive-Agentic-AI-into-Any-Workflow-Create-Multimodal-Experiences-and-Extend-Digital-Labor-Throughout-the-Enterprise/default.aspx>

- Salesforce. (2025j). Salesforce launches Agentforce 3 to solve the biggest blockers to scaling AI agents: Visibility and control. [Salesforce lanza Agentforce 3 para resolver los mayores obstáculos para escalar los agentes de IA: visibilidad y control.] <https://investor.salesforce.com/news/news-details/2025/Salesforce-Launches-Agentforce-3-to-Solve-the-Biggest-Blockers-to-Scaling-AI-Agents-Visibility-and-Control/default.aspx>
- Salesforce Help. (2025a). Agentforce Testing Center overview. [Descripción general del Centro de pruebas de Agentforce.] https://help.salesforce.com/s/articleView?id=ai.agent_testing_center.htm&type=5
- Salesforce Help. (2025b). Asignaciones de Salesforce Developer Edition. https://help.salesforce.com/s/articleView?id=xcloud.overview_limits_developer.htm&type=5
- Santander. (2025). Santander acelera su estrategia Data & AI-First tras firmar un acuerdo con OpenAI. <https://www.santander.com/es/stories/santander-acelera-su-estrategia-data-ai-first-tras-firmar-un-acuerdo-con-openai>
- República Argentina. (2000). Ley N° 25.326, Protección de los datos personales. <https://servicios.infoleg.gob.ar/infolegInternet/anexos/60000-64999/64790/texact.htm>
- Tavasoli, M., Ziegler, T., Wu, Y., & Yang, J. (2025). Responsible innovation: A strategic framework for financial LLM integration. [Innovación responsable: Un marco estratégico para la integración de LLM financiero.] <https://arxiv.org/abs/2504.02165>
- The Wall Street Journal. (2025). The Rise of Artificial Intelligence at JPMorgan. [El auge de la inteligencia artificial en JPMorgan.] <https://www.wsj.com/tech/ai/jpmorgan-chase-artificial-intelligence-banking-939b1b32>
- Todupunuri, A. (2024). Virtual Assistants and Chatbots: Exploring the Impact of AI-Powered Virtual Assistants on Customer Service Efficiency and Satisfaction. [Asistentes virtuales y chatbots: exploración del impacto de los asistentes virtuales impulsados por IA en la eficiencia y la satisfacción del servicio al cliente.] <https://irjems.org/Volume-3-Issue-10/IRJEMS-V3I10P124.pdf>

-
- Turing, A. M. (1950). Computing machinery and intelligence. [Maquinaria computacional e inteligencia.] <https://doi.org/10.1093/mind/LIX.236.433>
- Unión Europea. (2024). Reglamento (UE) 2024/1689 sobre inteligencia artificial (AI Act) [Reglamento de la UE sobre inteligencia artificial] https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=OJ:L_202401689
- Valoir (2025). Accelerating time to agentic AI value. [Acelerando el tiempo para obtener valor de la IA agente.] <https://www.salesforce.com/en-us/wp-content/uploads/sites/4/documents/research/valoir-report-accelerating-agentic-ai-time-to-value.pdf>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. [La atención es todo lo que necesitas.] <https://arxiv.org/abs/1706.03762>
- World Economic Forum & Accenture. (2024). The future of financial advice. [El futuro del asesoramiento financiero.] https://www3.weforum.org/docs/WEF_The_Future_of_Financial_Advice_2024.pdf
- Wu, S., Irsoy, O., Lu, S., Dabrovolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). BloombergGPT: A large language model for finance. [BloombergGPT: Un gran modelo de lenguaje para las finanzas.] <https://arxiv.org/pdf/2303.17564v3>
- Yang, H., Liu, X.-Y., & Wang, C. D. (2023). FinGPT: Open-Source Financial Large Language Models. [FinGPT: Modelos de lenguaje financiero de código abierto.] <https://arxiv.org/abs/2306.06031>
- Yang, Y., Siy UY, M. C., & Huang, A. (2020). FinBERT: A pretrained language model for financial communications. [FinBERT: Un modelo de lenguaje preentrenado para comunicaciones financieras.] <https://arxiv.org/abs/2006.08097>
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Ju, X., Chen, W., Zhang, P., Zhang, Y., Dong, Z., Yin, D., & Wen, J.-R. (2023). A survey of large language models [Una revisión de los modelos de lenguaje a gran escala] <https://arxiv.org/abs/2303.18223>