Universidad de los Andes

Facultad de Economía

METODOLOGÍA, Y MODELO FINANCIERO SOBRE LA IMPLEMENTACIÓN DE MODELOS DE LENGUAJE DE GRAN TAMAÑO PARA CHATBOTS EN SERVICIO AL CLIENTE EN EL MERCADO COLOMBIANO

Asesores: Rubén Manrique/Leonardo García

Presentado por: Julián David Ríos López, (201517163)

20 de octubre de 2024

Resumen

En la actualidad, la calidad del servicio al cliente se ha convertido en una prioridad para las empresas, ya que impacta directamente en la captación y retención de usuarios. Conscientes de ello, y gracias al avance de la inteligencia artificial, muchas organizaciones han implementado herramientas basadas en *machine learning* para mejorar la atención al cliente a través de una mayor cobertura y disponibilidad a través de sus diferentes canales digitales.

En este contexto, la siguiente investigación se centra en estudiar diferentes aproximaciones de modelos de lenguaje de gran tamaño (LLM) dirigidas a servicio al cliente, seleccionando y desarrollando un modelo de generación aumentada por recuperación (RAG). El objetivo es crear una metodología que ayude a las empresas de diversos sectores a implementar, ya sea de forma nativa o comercial, estos modelos en su esquema servicio al cliente. Dicha metodología se acompaña de un modelo financiero y un análisis cualitativo que permite responder a la pregunta de cuáles son los costos asociados a la implementación de dichos modelos en servicio al cliente en el mercado colombiano, y cuales características implícitas pueden llevar a una mejorar en las capacidades de este (en términos de efectividad, cobertura y disponibilidad). Así mismo, dicho modelo ofrece una herramienta a las empresas para escoger entre dos diferentes aproximaciones, una implementación nativa y una implementación con modelos comerciales, que según las características de cada negocio se puedan adaptar mejor a sus necesidades.

Los resultados de la investigación revelan una disminución en los costos tanto con la implementación nativa como con la contratación de modelos comerciales (incluido su desarrollo) para un caso base estándar de una empresa de mediano tamaño en el mercado colombiano comparado con los modelos de atención al cliente tradicionales por medio de call center, acompañado de una mayor cobertura, disponibilidad y obtención de datos sobre los usuarios dirigidos a una mejora cualitativa del servicio. Lo anterior muestra el potencial que tiene dicha tecnología de producir un impacto financiero positivo en las empresas, al llevar a un uso eficiente de los recursos, acompañado de una mejora en la capacidad analítica de los usuarios que tienen el potencial de generar una mejor calidad en el servicio al cliente. Así mismo, los resultados de este estudio, acompañado de los debidos supuestos mencionados, permitirán a las empresas interesadas de diferentes sectores hacer una

simulación de los costos asociados a dichas implementaciones basados en sus propias características y necesidades.

Palabras clave: Servicio al cliente, chatbots, inteligencia artificial, modelos de lenguaje de gran tamaño, generación aumentada por recuperación, análisis financiero.

Introducción

El servicio al cliente se ha convertido en una herramienta fundamental para fidelizar a los usuarios en diversos sectores, siendo un factor determinante para estos en su elección entre empresas. Según una encuesta realizada por Salesforce (2020) a más de 15000 usuarios, el 89% de los consumidores está más dispuesto a realizar otra compra a una marca que haya brindado un buen servicio al cliente. Así mismo, una investigación de HubSpot (2018) encontró que el 68% de los usuarios está dispuesto a pagar más por un producto de una marca con buena reputación en cuanto a servicio al cliente.

Por otro lado, los avances en materia de inteligencia artificial han llevado a las empresas a adoptar diferentes herramientas basadas en esta que les permitan aumentar su capacidad en diferentes aspectos de sus productos incluido el servicio al cliente, como por ejemplo el uso de chatbots, con el objetivo de ampliar su cobertura, así como su disponibilidad. Un estudio realizado por Tiny Talk (2024) encuentra que empresas como Amazon, Marriot, o Walmart han implementado chatbots para el soporte de sus clientes, obteniendo beneficios como acceso instantáneo de atención a sus usuarios, interacciones personalizadas, entre otros.

Sin embargo, a pesar del crecimiento en la investigación sobre modelos de lenguaje de gran tamaño (LLM)[1], los cuales son utilizados para la implementación de dichos chatbots, en la actualidad existe poca literatura relacionada a dos partes esenciales dentro de este proceso:

1. **Metodología de implementación:** se necesita una metodología robusta que permita implementar dichos procesos, ya sea de manera nativa o comercial (ver Apéndice 1

[1] Los modelos de lenguaje de gran tamaño son sistemas de inteligencia artificial con la capacidad de procesar y generar texto en grandes cantidades.

para ahondar en estas definiciones), dentro de una organización que tenga en cuenta las diferentes necesidades de estas (por ejemplo, en términos de cobertura, disponibilidad y calidad de las respuestas), y que les ofrezca unos lineamientos y recomendaciones que permitan una adopción fluida y eficiente de estas tecnologías.

2. **Análisis comparativo con soluciones existentes:** es necesario realizar un análisis comparativo en términos de costos y capacidades de dichas implementaciones versus soluciones tradicionales como call center, lo cual permitiría a las empresas tomar mejores decisiones según sus necesidades.

Basado en lo anterior, este trabajo responde a la pregunta: ¿la implementación de LLM, tras el desarrollo de una metodología de implementación nativa y comercial, conlleva a una mejora en las capacidades del servicio al cliente y una reducción de costos? Los resultados muestran que existe una disminución en los gastos asociados a este rubro y que existen ventajas que no presentan sistemas de atención tradicionales como la recolección de datos de los usuarios para una mejor capacidad analítica, así como un aumento en la escalabilidad, disponibilidad, y cobertura, lo cual se puede traducir en una mejor calidad en el servicio. Estos hallazgos evidencian el potencial financiero que representa la implementación de dichos modelos para las empresas, tanto en términos de uso eficiente de los recursos como en la fidelización de usuarios por medio de un mejor servicio.

Revisión de Literatura

Taxonomía	
Dimensión	**Trabajos**
Elección del modelo	He et. al. (2023)
	Ovadia et. al. (2023)
Implementación de modelos de lenguaje de gran tamaño en chatbots	Singhal et. al. (2023)
	Li et. al. (2023)
Desarrollo de metodología e implementación	Es et. al. (2023)

Efectividad de LLM en servicio al cliente	Camilleri y Troise (2023)
	Hwgan y Kim (2021)
	LocaliQ(2023)
Aspectos que se deben tener en cuenta para la interacción chatbot-persona	Roller et. al. (2020)
	White, Marating y White (2022)

Tabla 1 Taxonomía de la revisión de literatura

El aporte de este trabajo se encuentra en investigar y desarrollar una metodología de implementación nativa de LLM, junto con un análisis financiero y de calidad que permita a las empresas determinar si este tipo de soluciones se ajustan a las necesidades de su negocio.

Por un lado, existe literatura que respalda el uso de LLM para la creación de chatbots. He et. al. (2023) en un trabajo de investigación reciente exploran las características actuales que permiten una transición desde los modelos de lenguaje preentrenados *(*PLM) utilizados anteriormente para este fin, hacia LLM, incluyendo el tamaño de las redes neurales empleadas en dichos modelos, el volumen de datos existentes en la actualidad, y una mayor capacidad de cómputo. Sus hallazgos sugieren que los LLM superan a los PLM en medidas de desempeño como la precisión de las respuestas. Por ende, el estado actual de la investigación en este campo respalda el uso de LLM para la creación de herramientas como chatbots.

En cuanto a la elección de una implementación de LLM que permita tener en cuenta el contexto de negocio necesario dentro de la metodología que se desarrolla en este trabajo, la investigación sugiere que la generación mejorada por recuperación (RAG)[2] es la opción indicada para dicha tarea. Ovadia et al. (2023) concluyen que la implementación de RAG presenta una ventaja significante comparado a implementaciones de LLM por medio de *fine-*

[2] La generación aumentada por recuperación (RAG) es una técnica que busca mejorar la calidad y la precisión de la generación de texto por medio de la recuperación de información desde una base de conocimiento, utilizando esta como contexto a la hora de responder una pregunta por parte del usuario.

tuning[3], principalmente porque no solo enriquece el modelo con nuevo conocimiento, sino que también incorpora contexto relevante a la pregunta que se está respondiendo. Así mismo, Hernández (comunicación personal, abril 1 de 2024) afirma que desde su experiencia una implementación por RAG incurre en menos costo de procesamiento de información y entrenamiento del modelo que *fine-tuning*. Por lo tanto, se puede observar que la elección de RAG presenta ventajas tanto en términos de calidad de las respuestas entregadas como de la calidad de los costos.

Así mismo, existen diversas investigaciones que implementan LLM para el desarrollado de chatbot, las cuales sirven como guía para la metodología que se desarrolla en este trabajo. Singhal et. al. (2023) desarrollan Med-PaLM2, un modelo que mejora en un 19.3% la precisión (de 67.2% a 86.5%) respecto a chatbots dirigidos al sector salud. Para ello, utilizan técnicas como prompt de pocas muestras[4], autoconsistencia[5] y refinamiento de ensamblaje[6]. Por otro lado, Li et al (2023) crean ChatDoctor, una herramienta basada en Llama, un LLM de código abierto, para atender dudas sobre salud. Esto muestra que el desarrollo de estas tecnologías no depende de modelos comerciales, y que son accesibles al público. Estos ejemplos evidencian la variedad de factores que influyen en la creación de una metodología para implementar modelos LLM, dependiendo su objetivo: en el caso de esta investigación, enfocado a servicio al cliente.

En cuanto al desarrollo de la metodología, son varias investigaciones las que sugieren unos marcos de trabajo que permitan una elección correcta de los diferentes parámetros que se deben tener en cuenta a la hora del desarrollo de esta. Por un lado. Lewis, P. et. al. (2021) en la investigación que da origen a RAG, muestran las ventajas que tiene este tipo de implementación, logrando superar el estado del arte en tareas relacionadas con Pregunta/Respuesta (QA). Por otro lado, para el proceso de evaluación, Es et. al. (2023) desarrollan un marco de trabajo que permite evaluar arquitecturas de desarrollo que

[3] El *fine-tuning* es una técnica utilizada para reentrenar principalmente LLM con información de un rubro en específico, con el objetivo de ajustar el conocimiento del modelo a dicho dominio.

[4] Técnica basada en el concepto de cadena de pensamiento que permite mantener el contexto de una conversación.

[5] Técnica que consiste en tener una muestra de múltiples explicaciones y respuestas al texto ingresado por el usuario provenientes de la base de conocimiento.

[6] Técnica que consiste en generar respuestas a lo escrito por el usuario de manera estocástica, para luego ingresarlas dentro del *prompt* original con el objetivo de generar una respuesta final.

implementan RAG, en donde implementan métricas que permiten iterar entre diferentes alternativas de cada uno de los parámetros de este marco de trabajo (ver Apéndice 1). Teniendo en cuenta estas herramientas, se puede generar una metodología que ofrezca valor a las empresas que quieran implementar sistemas de atención al cliente manejados por medio de inteligencia artificial.

Adicionalmente, existen investigaciones que exploran las posibles implicaciones que pueden tener la implementación de LLM dirigidos al servicio al cliente en diferentes industrias. Camilleri y Troise (2023) encuentran en su investigación que solo el 15% de las solicitudes que llegan a servicio al cliente necesitan de atención humana, y que los costos potenciales de dichas implementaciones, y en los que hay que prestar atención, están relacionados con factores cognitivos, de afectividad, y de funcionalidad. Sin embargo, no ahondan en las implicaciones financieras de dichas implementaciones, comparadas con otros canales de servicio al cliente tradicionales. Por otro lado, Hwang y Kim (2021) investigan el impacto de LLM en el sector financiero en el área del servicio al cliente y encuentran que no existen diferencias significativas de ingresos recibidos por instituciones financieras entre usuarios que utilizan medios tradicionales y chatbots para productos de uso establecido, aunque si hay diferencias negativas frente a interacciones relacionadas con nuevos productos. Adicionalmente, según una investigación realizada por Malino (2023), los negocios que proveen servicios de chatbot a sus clientes experimentan un aumento del 70% en las interacciones que sus clientes tienen con este. Estos hallazgos sugieren que la implementación de LLM en la atención al cliente puede ofrecer ventajas, pero también subrayan la importancia de definir y gestionar lo diferentes riegos asociados.

Finalmente, se encontraron investigaciones que guían los parámetros cualitativos que se deben tener en cuenta a la hora de desarrollar herramientas como chatbots para que estos presenten unas características que hagan deseable su interacción con personas. Roller et. al. (2020) encuentran en su investigación que existen características de modelo de chatbot que se deben tener en cuenta para mejorar la interacción con las personas, entre los que se incluye empatía con el usuario, proveer fuentes de conocimiento, entre otras. En su investigación concluyen que los resultados de un modelo que es tuneado teniendo en cuenta estos puntos presenta una mejor evaluación entre humanos que modelos con mayor número de parámetros. De la misma manera, White, Marating y White (2022) encuentran en una revisión de literatura

a chatbots diseñados durante la pandemia del COVID-19 que existen características facilitan la adopción de estos sistemas como son la relevancia del contenido, la confianza que se tiene en la herramienta, la habilidad digital de los usuarios, acompañado de la aceptabilidad que tengan los usuarios. Se puede entonces que muchas de las características cualitativas que se buscan a la hora de implementar sistemas de atención al cliente con inteligencia artificial están relacionadas con características intrínsecas de una interacción humana

Metodología

La metodología desarrollada dentro del marco de este trabajo consiste en los siguientes pasos:

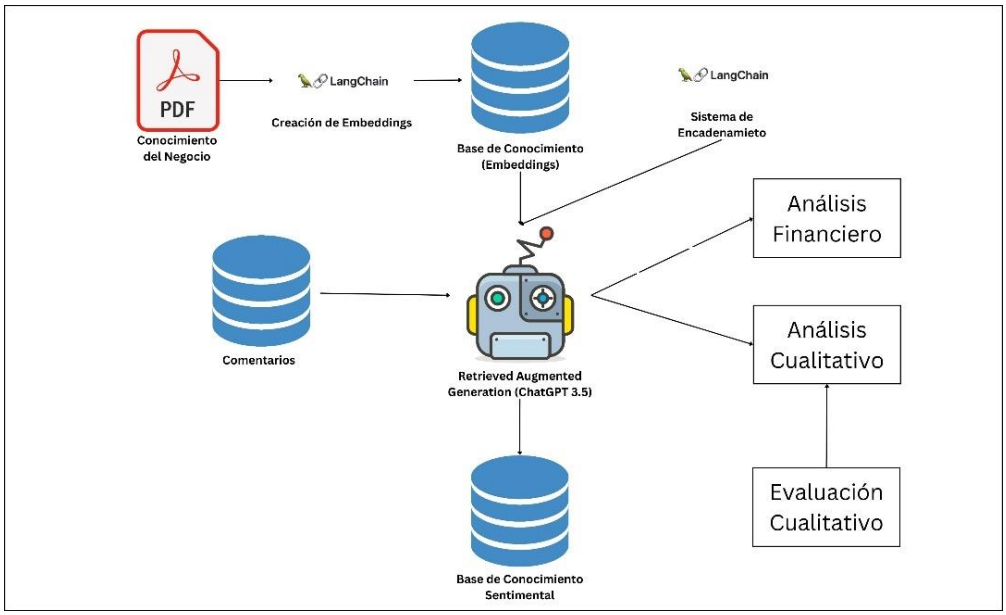

Ilustración 1: Metodología de implementación

a. Elección del marco de trabajo del LLM

En la revisión de literatura se encontró que el marco de trabajo más adecuado para desarrollar un chatbot de servicio al cliente basado en una implementación de un LLM es el RAG. Este marco permite una mejora del modelo base, por medio de una reducción en las alucinaciones[7] y la combinación de diferentes fuentes de conocimiento, así como la reducción de costos frente a otras alternativas como el *fine-tuning*. Así mismo, para la

[7] Las alucinaciones son respuestas que la IA da como ciertas basadas en datos que no están dentro de su entrenamiento.

implementación de esta metodología dentro de la investigación se utilizará el modelo de ChatGPT 3.5, basado en la API de Open AI, lo anterior debido a su robustes de entrenamiento (175MM de parámetros) y sus costos frente a modelos como ChatGPT 4.

b. Procesamiento y almacenamiento de los datos

Para el almacenamiento de los datos en una base de conocimiento necesaria para implementar RAG, es necesario procesar los datos para convertirlos en *embeddings*[8], con el objetivo de poder utilizar posteriormente dichas representaciones en una búsqueda de similaridad entre la base de conocimiento y lo ingresado por el usuario. Así mismo, para construir los *embeddings*, es necesario dividir los textos que se desean almacenar en partes de un tamaño de *tokens*[9] determinado, llamadas *chunks*. La elección del número de *tokens* por *chunk*, así como el *overlapping*[10] de estos depende del caso de uso, así como del modelo que se elige y su ventana de contexto[11] (en el caso de ChatGPT 3.5 es de 4096 tokens). En el caso de esta investigación, para realizar los *embeddings* se utiliza la API de OpenAI usando el modelo "text-embedding-ada-002". Finalmente, dichos datos se almacenan en bases de datos diseñadas para representaciones vectoriales de este tipo, de las cuales para el caso de uso de esta investigación se usó Pinecone[12].

Adicionalmente, para el análisis sentimental se procesa los textos ingresados por el usuario con el fin de determinar sentimiento de la persona acerca del servicio (i.e. positivo, neutro, negativo). Dicho análisis permite identificar datos propios del lenguaje natural de la opinión de manera automática (cosa que se encontró en la revisión de literatura es relevante para una mejora constante en el servicio). Finalmente se almacena el costo que las interacciones tienen (solamente aplica para el modelo comercial). Estos datos se almacenan en bases de datos relacionales estándar.

c. Diseño del prompt

---

[8] Representaciones vectoriales de palabras o frases, con el fin de codificar el significado de una frase.
[9] Son unidades básicas de texto que se utilizan para el procesamiento de lenguaje natural. En el caso de ChatGPT representa 4 caracteres en inglés.
[10] Es una estrategia donde *chunks* consecutivos comparte algunos *tokens.*
[11] La ventana de contexto es la cantidad de información que se puede registrar dentro del *prompt* de una modelo.
[12] Existen otras interesantes en el mercado como Qdrant y Chroma.

Para la comunicación con los modelos LLM que van a trabajar con RAG, es necesario diseñar un *prompt*[13] que permita interactuar con estos, de manera que tenga en cuenta las recomendaciones encontradas en la revisión de literatura a la hora de implementar chatbots que interactúan con personas. Teniendo en cuenta lo anterior, el *prompt* diseñado en esta metodología tiene en cuenta los siguientes componentes:

- Pregunta: la cual consiste en el input que introduce el usuario por medio de la interfaz gráfica.

- Contexto: se trata de los documentas que encuentra la búsqueda de similaridad que se ajustan a la pregunta ingresada por el usuario.

- Historia del chat: la cual recoge las interacciones entre el usuario y el modelo, con el objetivo de almacenar información relevante de la conversación (e.g. nombre del usuario).

- Sentimiento: para que el lenguaje del modelo se adapte a las necesidades del usuario y sienta más cercanía con el modelo (a diferencia de si sintiera que está hablando con una maquina), el prompt tiene en cuenta el sentimiento del usuario dentro de la interacción. Para capturar dicha información, se utiliza la librería *pysentimiento*.

d. Desarrollo del algoritmo

El algoritmo consiste en recibir la pregunta por parte del usuario. Después de esto, el chatbot analiza la información acerca del sentimiento que tiene el comentario (positivo, negativo y neutro), y teniendo en cuenta este devuelve una respuesta generada por el contexto dado por los *embeddings* que más se ajusten al texto ingresado. Posterior a esto, la interacción continua, dejando en la historia del chat las interacciones previas a manera de memoria. A medida que se desarrolla la conversación, se va modificando tanto el sentimiento del usuario, como los costos en los que se incurre debido a las interacciones (solo para la implementación con modelos comerciales). Dicha información se almacena en una base de datos para su posterior análisis.

[13] Instrucciones que reciben el modelo.

e. Evaluación del modelo y perfeccionamiento

Se realiza una evaluación técnica del modelo en donde se busca determinar la precisión que tiene este en cuanto a las respuestas que se buscan, con el objetivo de realizar un posterior perfeccionamiento de este (i.e. elección de parámetros, pre y post procesamiento de la información, entre otros). Dicha evaluación se realiza por medio de RAGAS[14], un marco de trabajo que permite la evaluación de pipelines que implementan RAG (Es et. al., 2023). Este tiene en cuenta 4 aspectos a la hora de evaluar las respuestas entregadas por el modelo: fidelidad, precisión del contexto, relevancia de la respuesta, y recuperación del contexto (ver Apéndice 2).

f. Análisis financiero

Finalmente, se realiza un análisis financiero que determine el comportamiento a nivel de costos de la implementación de las diferentes alternativas, para determinar si es acorde a los presupuestos en este rubro, así como realizar una comparación que pueda determinar si existe una reducción de costos, acompañada de una expansión en la cobertura y disponibilidad del servicio, lo cual desemboque en una mejora en el servicio prestado por estas empresas en general, sin afectar la calidad esperada de este. Para el caso de este trabajo, se utilizará una proyección de costos basado en la información recolectada del funcionamiento de un call center por 3 meses, junto con toda la información mencionada en la descripción de los datos, por medio de un modelo financiero que las organizaciones interesadas podrán tener a su disposición como resultado de esta investigación.

Descripción de los Datos

Los siguientes son los datos que se recolectaron para la realización del modelo financiero:

a. Caso de estudio call center

Para el desarrollo del caso base de estudio se tomó una muestra de un call center que tiene las siguientes características:

Información Call Center Mensual	
Número de usuarios activos	594000

[14] *Retrival Augmented Generation Assesment.*

Usuarios que realizan algún contacto (en promedio)	4520
Ratio de usuarios que contactan	0.008
Número de contactos/sesiones (en promedio)[15]	9338
Minutos de diálogo (en promedio)	23144
Número de empleados	9

Tabla 2 Datos call center

Así mismo, se obtuvieron datos financieros de proveedores de los insumos necesarios para call center autogestionados (i.e. licencias, líneas y minutos), y se escogió para el caso de estudio el que presenta menores costos (Five9) entre los consultados (Five9, Zendesk, e inConcert), y teniendo en cuenta que presentan las características técnicas similares.

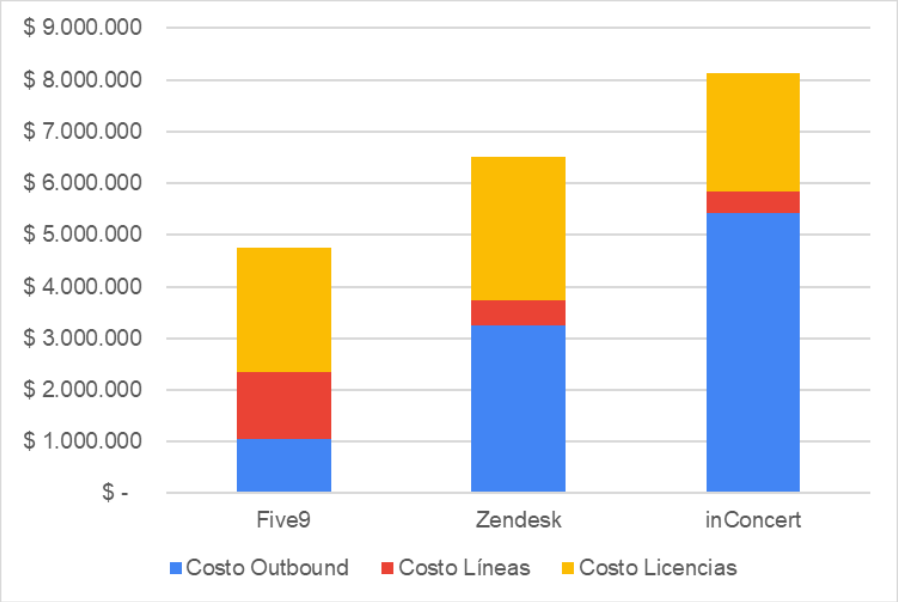

Ilustración 2 Comparación costos proveedores call center

b. Caso de estudio modelo comercial

[15] Hace referencia a las veces que un usuario se ha puesto en contacto con servicio al cliente.

Para el desarrollo del modelo financiero que consume modelos comerciales se tuvieron en cuenta los siguientes modelos:

- ChatGPT 4 Turbo: el cual se utiliza como modelo conversacional, y que cuenta con una ventana de contexto de 128K tokens, con unos costos de USD30 por cada millón de tokens para el input y de USD60 para el output.
- ChatGPT 3.5 Turbo: el cual se utiliza como modelo conversacional, y que cuenta con una ventana de contexto de 128K tokens, con unos costos de USD30 por cada millón de tokens para el input y de USD60 para el output.
- ada v2: el cual se utiliza para realizar los embeddings del texto, y cuesta USD0.10 por cada millón de tokens que procesa.

c. Caso de estudio modelo nativo

Para el desarrollo del modelo financiero que sugiere una implementación nativa se tuvieron en cuenta los siguientes modelos:

- zephyr-7b-beta: desarrollado por Tunstall et. al. (2023), se utiliza como modelo conversacional de código abierto. Este modelo se escogió ya que es el modelo de 7B de parámetros que cuenta con el ranking más alto los rankings de MT-Bench y Alpaca Eval, así como que permite interactuar por medio de la API de Hugging Face.
- bge-small-en-v.1.5: desarrollado por Xiao, Liu y Muennighoff (2023), se utiliza como modelo de código abierto para realizar los *embeddings*. Dicho modelo se encuentra como estado del arte de modelos de embeddings de código abierto.
- Costos: para la implementación de este tipo de modelos, es necesario montarlos en una tarjeta gráfica que soporte el modelo. Para esta investigación, se utiliza una GPU NVIDIA Tesla P40 montada en AWS. El costo por hora de esta tarjeta es de USD 4.88.

d. Supuestos del modelo

Para la construcción del modelo financiero, se combinaron los hallazgos de la revisión de literatura con los datos obtenidos de entrevistas tanto a personal administrativo como técnico, definiendo así los siguientes supuestos:

- Dado que el escenario base se plantea con el objetivo de poder cambiar las metodologías tradicionales de call center, se asume que la inversión inicial en este caso es de 0.
- Por la investigación realizada previamente, se supone que una implementación de modelos de chatbot para atención al cliente trae consigo un incremento en los usuarios que contactan, por lo cual en el escenario tradicional de call center no se tiene en cuenta.
- El caso base estudia un modelo de startup, en el cual se espera un crecimiento esperado anual del 10% en los usuarios de manera natural (Duque, comunicación personal, septiembre 23 de 2024). Este crecimiento para simplicidad del modelo no afecta el número de agentes en el call center, pero sí afecta los minutos de dialogo mensual en la misma proporción.
- Al investigar el comportamiento histórico que han tenido los precios de los diferentes servicios de tecnología se asume:
    - Un incremento de 20% en los costos de infraestructura de call center.
    - Un incremento en el almacenamiento en la nube de 10.60%.
    - Una disminución en los costos de procesamiento en la nube para modelos nativos de 15%.
    - Un aumento en los costos de la API de OpenAI de 5%.
- Al indagar con personal técnico (Hernández, comunicación personal, abril 22 de 2024) acerca de los costos de desarrollo que tenían tanto la implementación nativa como comercial se encontró que:
    - Las horas de desarrollo de una implementación nativa son cerca de 20.
    - Las horas de desarrollo de una implementación comercial son cerca de 5.
    - El costo hora de un desarrollador para esta tarea es USD75
- Adicionalmente, se tiene aspectos del mercado colombiano como:
    - Un aumento promedio del salario del 10%.
    - Una tasa interbancaria a la fecha del 12.22%.
    - Un precio del dólar de 4100 pesos colombianos.
- Se agregó un parámetro “beta” que permite incrementar o disminuir el número de interacciones base que tiene una sesión promedio por usuario (la cual está fijada en 5), y el cual permite ajustar el modelo a diferentes casos de uso.

- Se agregó un parámetro “alfa” que tiene en cuenta el *benchmark* del MMLU[16], y que permite aproximar la exactitud que tendrán los modelos a la hora de responder los requerimientos de los usuarios, y que permite incrementar el número de interacciones necesarias para poder resolver una petición. El valor para cada modelo es de:
    - **ChatGPT4:** 1
    - **ChatGPT3.5:** 1,23
    - **Zhepyr7B:** 1,41

Resultados

a. Modelo Chatbot

Como resultado de la investigación se desarrolló un modelo que permite interactuar por medio de una interfaz gráfica con un asistente virtual, el cual carga información desde la base de conocimiento, permitiendo traer información de interés acerca de la petición del usuario relacionada con el negocio de interés, con el fin de responder las inquietudes del usuario. El modelo que alimenta dichas respuestas puede variar según la implementación deseada.

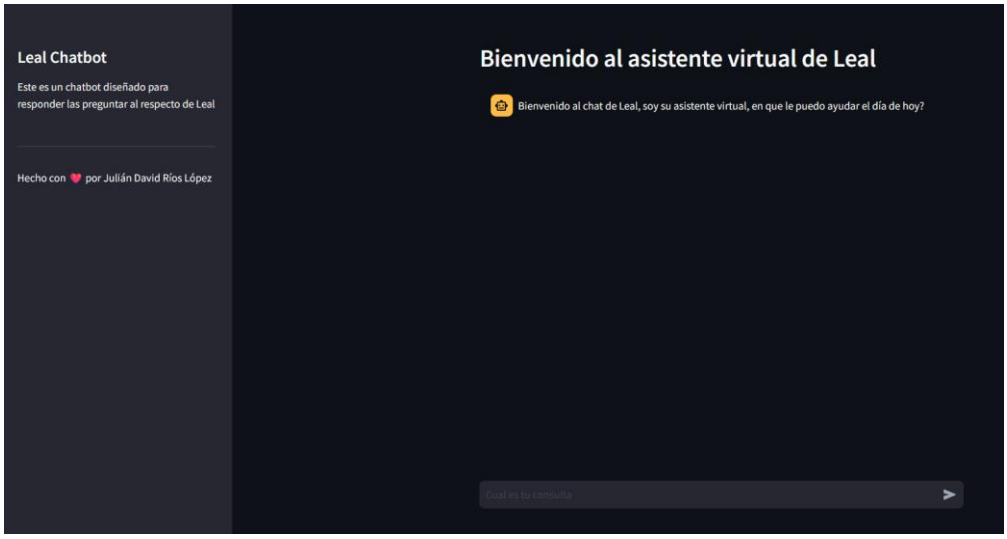

Ilustración 3 Interfaz gráfica chatbot

[16] El MMLU (Masive Multitasking Language Understanding) es una métrica que evalúa modelos de lenguaje en tareas de comprensión y razonamiento, y la cual se utiliza como un estándar para medir el rendimiento de modelos de procesamiento de lenguaje natural.

Al ser una aplicación que se puede implementar tanto de manera nativa como comercial, permite una escalabilidad tanto en términos de cobertura, así como de disponibilidad (24/7).

b. Comparación de los modelos

Basado en información de diferentes fuentes (Towards AI (2024), SourceForge (2024) y ZephyrLogic (2024)), se realizó la siguiente tabla que permite comparar los modelos:

Modelo	ChatGPT4	ChatGPT3.5	Zephyr7B
Parámetros	~1 trillón (americanos).	~175 billones (americanos).	~7 billones (americanos).
Multimodalidad	Sí (soporta entradas de texto e imagen).	No (solo soporta texto).	No (solo texto).
Rendimiento	Razonamiento y solución de problemas superiores.	Fuerte pero limitado en tareas más complejas.	Competitivo en generación de texto básica.
Multilingüe	Alta (maneja idiomas menos comunes).	Decente en idiomas principales.	Moderada.
Seguridad	82% menos propenso a producir contenido dañino (se puede mejorar con prompt).	Más propenso a respuestas sesgadas (se puede mejorar con prompt).	No se encontraron métricas de seguridad (se puede mejorar con prompt).
Alucinaciones	Mejoradas, pero aún presentes (se pueden mejorar con RAG)	Más frecuentes (se pueden mejorar con RAG).	Menos propenso a alucinaciones (se pueden mejorar con RAG).
Costo	Mayor debido a más poder computacional (costos tenidos en cuenta en modelo financiero).	Más económico que ChatGPT4.	Muy rentable, funciona incluso en computadores.
MMLU	86.4%	70%	61%

c. Mejoras cualitativas

Como resultado de la implementación de la metodología, se desarrolló una serie de dashboards que se alimentan con información que se genera automáticamente desde la

aplicación, principalmente del sentimiento de las personas (ver Ilustración 4). Lo anterior, según Duque (comunicación personal, abril 22 de 2024), quien lleva trabajando en servicio al cliente por más de 10 años, se traduce en una mejora en la capacidad analítica de las empresas en cuanto al comportamiento de sus usuarios, ofreciendo una herramienta a la hora de mejorar la calidad del servicio.

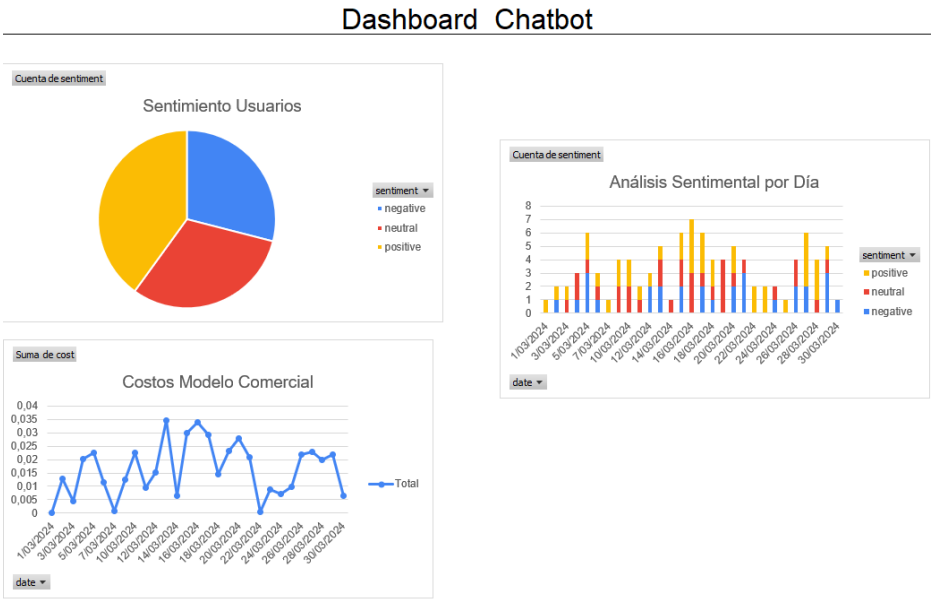

Ilustración 4 Dashboard de control

Adicionalmente, para el modelo comercial se recolecta información acerca de los costos que tiene cada contacto con el chatbot, para llevar así mismo un registro financiero acerca del uso de este.

d. Análisis financiero

Los resultados a la hora de evaluar el modelo financiero, bajo los supuestos del escenario base a dos años son los siguientes:

Análisis Financiero (2 años)	
Implementación	**VPN**
Call Center	-$ 492,008,897.90
GPT 4	-$ 116,499,540.49

GPT 3.5 Turbo	-$ 13,518,859.02
Nativa (Zhepyr7B)	-$ 311,671,633.77

Por otro lado, gracias al modelo financiero, se pudo calcular el tiempo en el que VPN de la implementación del modelo nativo sea indiferente a implementación comercial con ChatGPT4. Este punto es de interés ya que muestra que para las valuaciones que sean hechas con meses menores a este punto, la implementación con ChatGPT4 es menos costosa, recordando que esta implementación tiene unas capacidades mayores que las del modelo nativo del caso de estudio de esta investigación.

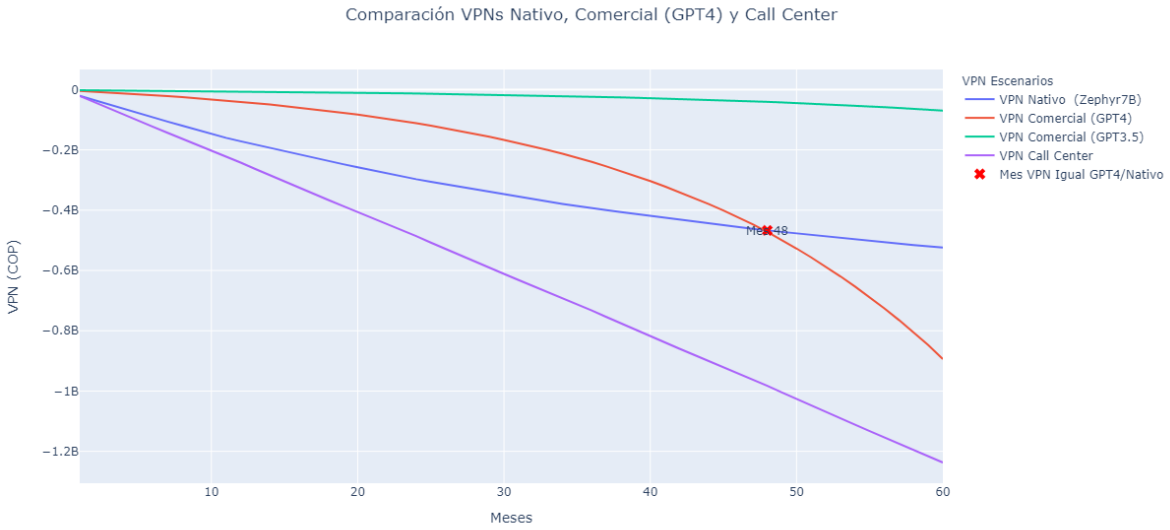

Ilustración 5 Comparación VPNs

Así mismo, el modelo permite encontrar cómo el comportamiento de variables exógenas, tales como el precio del dólar y el número de usuarios activos, puede afectar dicho punto, como se muestra en la ilustración 6 y 7.

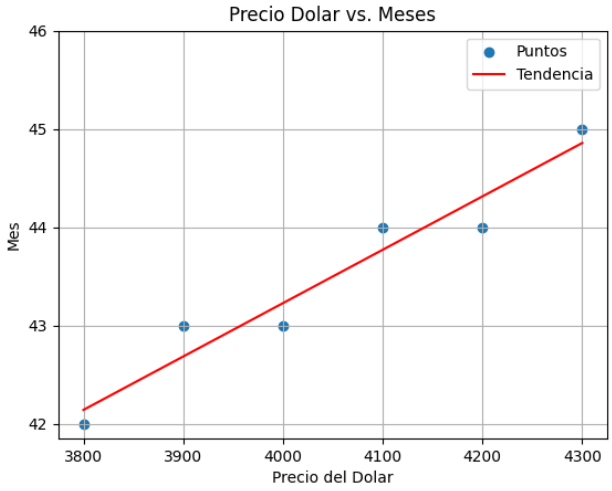

Ilustración 6 Cambio punto de cruce vs precio del dólar

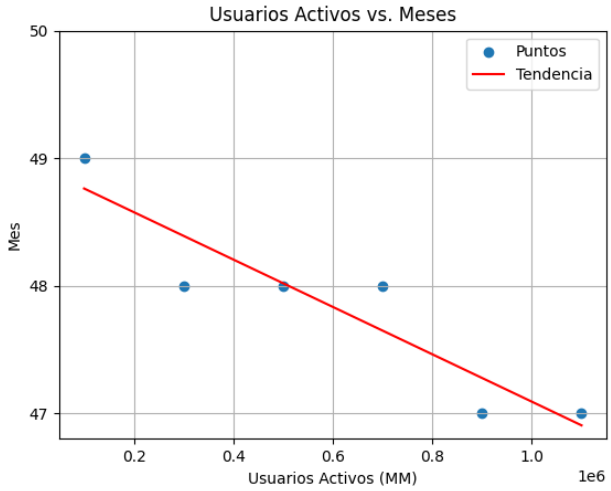

Ilustración 7 Cambio punto de cruce vs usuarios activos

Se puede observar que un aumento del precio de dólar hacer que este mes de equilibrio sea mayor, mientras que un aumento de los usuarios activos hace que disminuya.

Finalmente, como resultado de la investigación las empresas tendrán a su disposición el modelo financiero en donde podrán, basados en sus características propias, ver el comportamiento de los VPNs de las distintas implementaciones, así como los puntos de cruce que pueden ser de interés a la hora de tomar una decisión sobre cuál de estas elegir, como se muestra en la Ilustración 8.

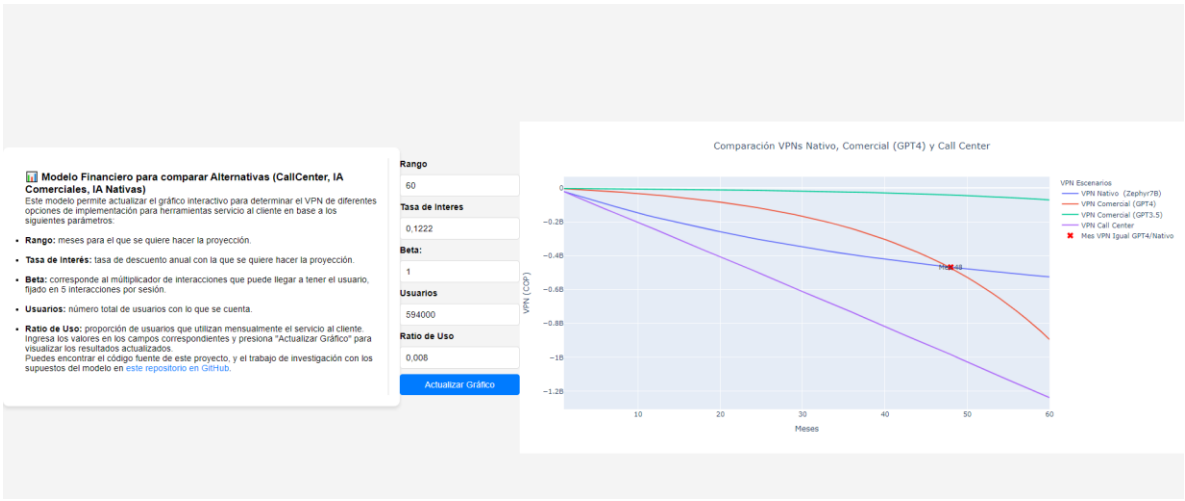

Ilustración 8 Modelo financiero disponible en el repositorio de la investigación

Análisis de Resultados

Los anteriores resultados se interpretan de la siguiente manera:

- El costo del call center presenta unos costos variables mensuales que dependen del número de agentes que se tengan en servicio. Lo anterior puede traducirse en una disminución en la capacidad de escalamiento de usuarios sin incurrir en costos.

- El costo promedio de una implementación que utilicen el modelo comercial de ChatGPT3.5 es casi un 90% menor que la que implementa ChatGPT4. Así mismo, la implementación por medio de RAG hace que la diferencia de la calidad entregada por los dos modelos no sea significante. Sin embargo, existen casos de uso en que es imprescindible la precisión del modelo, por lo que en este tipo de casos se sugiere la aproximación con ChatGPT4.

- El costo promedio de una implementación nativa, a pesar de ser alto, es incluso menor al modelo tradicional de call center. Así mismo, se puede observar que con el aumento de las interacciones producto de la adopción de estas tecnologías por parte de los usuarios, en algún punto a la hora de realizar la valuación de los proyectos se vuelve indiferente e incluso mejor opción escoger el modelo comercial. Es importante resaltar que existen actualmente en el mercado modelos más potente de libre acceso que pueden

igualar o mejorar la precisión de los modelos comerciales, pero que implican unos costos mayores de procesamiento.

Por otro lado, después de analizar los resultados encontrados, se puede llegar a describir las ventajas y desventajas de cada uno de los tres casos de estudio que se plantearon en este trabajo:

- **Modelo tradicional (call center)**
  - Ventajas: mayor cercanía con el usuario, mayor control en las respuestas.
  - Desventajas: mayores costos de escalabilidad, mayor riesgo de inconsistencias en la información (ya que existe cierta subjetividad propia de un ser humano).
- **Modelo comercial**
  - Ventajas: facilidad de implementación, disponibilidad hacia los usuarios.
  - Desventajas: costos de escalabilidad (entre más usuarios, mayores costos),
- **Modelo nativo**
  - Ventajas: eliminación de costos de escalabilidad, protección de datos.
  - Desventajas: costos de operación, principalmente de los procesos que corren en la nube.

Finalmente, se encuentra las siguientes limitaciones que se deben tener en cuenta:

Limitaciones del modelo

El planteamiento del modelo financiero cuenta con las siguientes limitaciones para tener en cuenta:

- Se está realizando una comparación entre una solución que ofrece una alternativa por medio de voz en tiempo real (call center) y otra que ofrece texto en tiempo real (chatbot). Lo anterior es debido a que las limitaciones que tienen los sistemas de voz potenciados por medio de LLMs tanto a nivel de latencia como de costos en la región.
- La comparación entre call centers y chatbots parte de la premisa de que ambas opciones podrían ofrecer una calidad de servicio similar. No obstante, debido a la imposibilidad de evaluar estas soluciones en un entorno real, no podemos se

puede afirmar con certeza que la calidad se mantenga. Sin embargo, basándose en la investigación previa, se pueden identificar una serie de ventajas significativas que los chatbots ofrecen. Estas ventajas sugieren que, en términos generales, podrían proporcionar una solución de calidad comparable. En particular, los chatbots destacan en áreas como la cobertura y disponibilidad, así como en la generación de datos valiosos que pueden utilizarse para mejorar la calidad del servicio. Estas características hacen que los chatbots no solo sean una alternativa viable, sino también una opción que potencialmente eleva la experiencia del cliente.

Conclusiones

Como se pudo observar a lo largo de la investigación, se pueden desarrollar sistemas de atención al cliente alimentados por inteligencia artificial que permiten aumentar la cobertura, mejorar la capacidad analítica sobre el comportamiento de los usuarios, y reducir costos frente a los modelos tradicionales de call center. Sin embargo, es necesario ahondar en aspectos cualitativos de estos modelos que permitan una mayor adopción por parte de los usuarios, reflejando características propias de una interacción humana como la empatía, y la confianza. Así mismo, utilizando herramientas de inteligencia artificial, se puede obtener información de las interacciones de manera automática (e.g. el sentimiento del usuario en la interacción), con el objetivo de encontrar datos relevantes acerca del comportamiento de los usuarios que no son evidentes a primera vista, principalmente en los requerimientos propios de servicio al cliente (PQRS), permitiendo de esta manera capturar información desde las inconformidades diarias de los usuarios del sistema. Adicionalmente, es necesario como siguiente paso en esta línea de investigación hacer una revisión *in situ* con organizaciones interesadas para evaluar la calidad que ofrecen estos modelos, medida de manera cuantitativa.

Con el objetivo de promover la adopción de la inteligencia artificial en el sector empresarial, este trabajo pone a disposición de los interesados las herramientas desarrolladas en el trascurso de la investigación (chatbot y modelo financiero), las cuales también permiten abrir nuevas vías de investigación, principalmente en temas relacionados al desarrollo de metodologías de implementación de sistemas de inteligencia artificial en diferentes rubros organizacionales, estudios que determinen el impacto a nivel financiero que pueden tener dichas implementaciones en las empresas, y el impacto que tiene la adopción de este tipo de

tecnologías en la economía en general, teniendo principalmente en cuenta el riesgo que se presenta en el mercado laboral con la adopción de estas tecnologías[17].

Agradecimientos

Quiero expresar mi más sincero agradecimiento a mis padres, Alfonso y Nubia, por el apoyo incondicional brindado a lo largo de toda mi carrera profesional. Agradezco profundamente a mis asesores, Rubén y Leonardo, por su valioso seguimiento y orientación durante la realización de este trabajo.

Asimismo, extiendo mi gratitud a Juan Sebastián y Cindy por su asesoría en temas administrativos y técnicos, y a Luis, cuyo apoyo desde la parte laboral fue de gran relevancia.

Finalmente, quiero reconocer a Santiago, Nicolás, Sebastián Ardila, Sebastián Cardona y Mallory por su constante acompañamiento y apoyo moral, sin los cuales este trabajo no hubiera sido posible.

Referencias

- Camilleri, M.A. and Troise, C. (2023). Chatbot recommender systems in tourism: A systematic review and a benefit-cost analysis. In Stockholm, Sweden: 8th International Conference on Machine Learning Technologies (ICMLT 2023), March 10–12, 2023, Woodstock, NY. ACM, New York, NY, USA, 10 páginas.
- Clarizia, F., Colace, F., Lombardi, M., Pascale, F., Santaniello, D. (2018). Chatbot: An Education Support System for Student. In: Castiglione, A., Pop, F., Ficco, M., Palmieri, F. (eds) Cyberspace Safety and Security. CSS 2018. Lecture Notes in Computer Science, vol 11161. Springer, Cham
- Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2023). RAGAS: Automated Evaluation of Retrieval Augmented Generation. arXiv.org.
- He. K et. Al. (2023). A Survey of Large Language Models for Healthcare: from Data, Technology, and Applications to Accountability and Ethics. Proceedings of the IEEE.

[17] El código se encuentra disponible en el siguiente repositorio: https://github.com/ViejoJuli/thesis_code

- HubSpot. (2018). Customer Service Expectation Survey. Consultado en: https://cdn2.hubspot.net/hubfs/2771217/Content/2018%20Customer%20Service%20Expectations%20Gladly.pdf
- Hwang, S. and Kim, J. (2021). Toward a Chatbot for Financial Sustainability. Sustainability 2021, 13, 3173.
- Li, Y. et. Al. (2023). ChatDoctor: A Medical Chat Model Fine-Tuned on a LLM Meta AI Using Medical Domain Knowledge. arXiv:2303.14070.
- Marino, S. (2023). 35+ Chatbot Statistics You Need to Know for 2024 | LocaliQ. Consultado en: https://www.localiq.com/blog/35-chatbot-statistics-you-need-to-know-for-2024/
- Rennie et. al. (2020). Scraping the Web for Public Health Gains: Ethical Considerations from a 'Big Data' Research Project on HIV and Incarceration. Consultado en: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7392638/
- Roller, S. et. al. (2020). Recipes for building an open-domain chatbot. Facebook AI Research.
- Salesforce. (2020). State of Connected Customer. Consultado en: https://c1.sfdcstatic.com/content/dam/web/en_us/www/documents/research/salesforce-state-of-the-connected-customer-4th-ed.pdf
- Singhal, K. et. Al. (2023). Towards Expert-Level Medical Questions Answering with Large Language Models. Google Research.
- Tiny Talk. (2024). 20 Companies Effectively Using AI Chatbots for Customer Support. Consultado en: https://www.linkedin.com/pulse/20-companies-effectively-using-ai-chatbots-customer-support-tiny-talk-p5j3e/
- Tunstall, L. et. al. (2023). Zephyr: Direct Distillation of LM Alignment.
- White BK, Martin, A & White, JA. (2022) User Experience of COVID-19 Chatbots: Scoping Review. J Med Internet Res. 2022;24(12).
- Xiao, S., Liu, Z. & Muennighoff, N. (2023). C-Pack: Packaged Resources To Advance General Chinese Embedding.

Bibliografía:

- Appy Pie. (n.d.). Datasets and data preprocessing for LLM training. Appy Pie. Consultado en: https://www.appypie.com/blog/datasets-and-data-preprocessing-for-llm-training
- Chatfield-Rivas, M. (n.d.). Post-process reliability. LinkedIn. Consultado en: https://www.linkedin.com/pulse/post-process-reliability-marie-chatfield-rivas
- Gal, P. (2023). Give chat agents personality by improving long-term memory systems. Medium. Consultado en: https://medium.com/@gal.peretz/give-chat-agents-personality-by-improving-long-term-memory-systems-daa7df20366c
- Hochmair, H. H., & Juhasz, L. (2024). Correctness Comparison of ChatGPT-4, Bard, Claude-2, and Copilot for Spatial Tasks. arXiv. Consultado en: https://arxiv.org/abs/2401.03506
- Labelerr. (n.d.). Data collection and preprocessing for large language models. Labellerr. Consultado en: https://www.labellerr.com/blog/data-collection-and-preprocessing-for-large-language-models/
- Peretz, G. (2023). Pre-processing and post-processing filtering systems for language models with help of sentiment analysis. ResearchGate. Consultado en: https://www.researchgate.net/publication/355446449_Pre-processing_and_Post-processing_Filtering_Systems_for_Language_Models_with_help_of_Sentiment_Analysis
- QueryVary. (n.d.). Optimizing LLM output: Step-by-step guide. QueryVary. Consultado en: https://www.queryvary.com/post/optimizing-llm-output-step-by-step-guide
- Scialom, T., & Bosselut, A. (2023). Data processing for large language models. Weights and Biases. Consultado en: https://wandb.ai/wandb_gen/llm-data-processing/reports/Processing-Data-for-Large-Language-Models--VmlldzozMDg4MTM2
- Unit8. (n.d.). A new era of AI: A practical guide to large language models. Unit8. Consultado en: https://unit8.com/resources/a-new-era-of-ai-a-practical-guide-to-large-language-models/
- Wang, A., et al. (2023). OpenAI's latest: A large-scale generative model [PDF]. arXiv. Consultado en: https://arxiv.org/abs/2310.05694

- Weidinger, L., et al. (2023). Advances in data preprocessing for AI [PDF]. arXiv. Consultado en: https://arxiv.org/abs/2312.05934
- Yan, E. (2023). Abstractive summarization: A review of methods. Eugene Yan. Consultado en: https://eugeneyan.com/writing/abstractive/
- Zhang, Y., & Li, T. (2024). Natural language understanding: A review. ACM Transactions on Intelligent Systems and Technology, 15(3), 35-58. https://doi.org/10.1145/3624062.3624172

Apéndices

1. Definiciones

- **Modelo nativo:** La implementación nativa en este contexto se refiere a aquellas soluciones que se ejecutan directamente en la infraestructura en la nube de una organización, aprovechando los *clusters* especializados en inteligencia artificial (IA) ofrecidos por proveedores como Amazon Web Services (AWS). Estos *clusters*, configurados con hardware y software optimizados para tareas de IA, permiten escalar los recursos de cómputo y almacenamiento de manera eficiente para entrenar y desplegar modelos de aprendizaje automático. En el caso de AWS, uno de los servicios más utilizados para estos fines es Amazon SageMaker, que proporciona un entorno completamente gestionado para construir, entrenar y desplegar modelos de machine learning a gran escala.
- **Modelo comercial:** La implementación comercial de modelos de inteligencia artificial hace referencia a el uso de APIs (*Application Programming Interfaces*) proporcionadas por diversos proveedores. Estos servicios ofrecen una amplia variedad de modelos preentrenados, cada uno especializado en tareas específicas. OpenAI, con su es uno de los más conocidos y utilizados en la industria. Sin embargo, existen otros proveedores como Google Cloud AI, Microsoft Azure AI y Hugging Face que ofrecen modelos y herramientas similares. La elección del proveedor y del modelo adecuado dependerá de las necesidades específicas de cada proyecto.

2. Métricas de RAGAS

- **Fidelidad:** mide la consistencia de la respuesta generado contra el contexto entregado en una escala de 0 a 1.

$$\boldsymbol{fidelidad} = \frac{|\boldsymbol{Número\ de\ afirmaciones\ en\ la\ respuesta\ que\ pueden\ ser\ inferidad\ del\ contexto}|}{|\boldsymbol{Número\ total\ de\ afirmaciones\ generadas}|}$$

- **Precisión del contexto:** evalúa si los elementos relevantes de la verdad fundamental presentes en el contexto son los mejores rankeados.

$$\boldsymbol{precion\ contexto@K} = \frac{\sum_{\boldsymbol{k=1}}^{\boldsymbol{K}}(\boldsymbol{Precision@k\ x\ v_k})}{\boldsymbol{Número\ total\ de\ items\ relevantes\ en\ el\ top\ K\ resultados}}$$

$$\boldsymbol{precion@k} = \frac{\boldsymbol{Verdaderos\ positivos\ \ @k}}{\boldsymbol{Verdaderos\ positivos\ @k + Falsos\ positivos\ @k})}$$

Donde *K* es el número total de chunk en el contexto y v_k (que es 0 o 1) es el indicador de relevancia el rango k.

- **Relevancia de la respuesta:** se enfoca en determinar cuan pertinente es la respuesta generada dado el *prompt*. Se ofrece un puntaje menor a aquellas respuestas incompletas o con contenido redundante.

$$\boldsymbol{relevancia\ respuesta} = \frac{\boldsymbol{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{cos}(\boldsymbol{E_{g_i}}, \boldsymbol{E_o})$$

Donde E_{g_i} es el *embedding* de la pregunta generada i, E_o es el *embedding* de la pregunta generado original, y *N* es el número de preguntas generadas (3 por defecto).

- **Recuperación del contexto:** mide hasta qué punto el contexto que ha sido recuperado se alinea con la respuesta dada, siendo está tomada como la verdad fundamental (VF).

$$\boldsymbol{recuperación\ contexto} = \frac{|\boldsymbol{Frases\ de\ la\ VF\ que\ pueden\ ser\ atribuidas\ al\ contexto}|}{|\boldsymbol{Número\ total\ de\ frases\ en\ la\ VF}|}$$