ECT | Revista Cientifica Emprendimiento Científico Tecnologico
ISSN: 2810 – 8493
https://revista.ectperu.org.pe/index.php/ect/index

Implementación de Chatbots con GPT como Asistentes Virtuales: Una Revisión Sistemática de la Literatura

Motta, B. [1*]; Salazar, A. [2]; Salcedo, A. [3]; Ticona, B. [4]

[1*] Facultad de Ingeniería de Sistemas e Informática, Universidad Nacional Mayor de San Marcos
https://orcid.org/0009-0003-9799-8160
brian.motta@unmsm.edu.pe

[2] Facultad de Ingeniería de Sistemas e Informática, Universidad Nacional Mayor de San Marcos
https://orcid.org/0009-0006-4517-5666
andhersson.salazar@unmsm.edu.pe

[3] Facultad de Ingeniería de Sistemas e Informática, Universidad Nacional Mayor de San Marcos
https://orcid.org/0009-0000-3859-9889
andres.salcedo@unmsm.edu.pe

[4] Facultad de Ingeniería de Sistemas e Informática, Universidad Nacional Mayor de San Marcos
https://orcid.org/0009-0000-4072-0641
arian.ticona@unmsm.edu.pe

Resumen:

Desde su aparición hace algunos años, los Grandes Modelos de Lenguaje (LLMs) han revolucionado el campo de los agentes conversacionales y asistentes virtuales. Conocidos comúnmente como chatbots, estos programas pueden ahora implementar alguno de los varios modelos de lenguaje disponibles y usar sus capacidades para realizar tareas específicas sin necesidad de realizar un trabajo exhaustivo. Con el continuo y rápido avance de modelos como LlaMA o GPT, ha aumentado también la cantidad de investigaciones para el desarrollo de chatbots en distintos campos. El presente trabajo busca hacer una revisión sistemática de la literatura de chatbots implementados con la familia de modelos GPT y responder a las preguntas de investigación referidas a 1) las tecnologías de uso más comunes en su implementación, 2) los problemas que conlleva este trabajo y 3) los métodos usados para evaluar el rendimiento del chatbot.

Palabras Claves: chatbot, GPT, asistente virtual, agente conversacional, asistente.

Abstract:

Since their appearance a few years ago, Large Language Models (LLMs) have revolutionized the field of conversational agents and virtual assistants. Known as chatbots, these programs can now implement any of the various language models available and use their capabilities to perform specific tasks without requiring extensive work. With the continuous and rapid advancement of models such as LlaMA or GPT, the amount of research for the development of chatbots in different fields has also increased. The present work seeks to make a systematic review of the literature on chatbots implemented with the GPT family of models and answer the research questions related to 1) the most common technologies used in their implementation, 2) the problems that this work entails and 3) the methods used to evaluate the performance of the chatbot.

Keywords: chatbot, GPT, virtual assistant, conversational agent, assistant.

1. Introducción

Para [1], la historia de los chatbots empieza en 1950, cuando Alan Turing planteó la pregunta: "¿Puede una máquina pensar?". A partir de esta reflexión, surgió la idea de que las máquinas podrían ser autónomas. Los chatbots, diseñados para responder preguntas automáticamente sin intervención humana, utilizan métodos de inteligencia artificial. Gracias a los avances recientes en este campo, los chatbots avanzados están reemplazando a los humanos en roles como soporte al cliente, asistentes de compras en línea y administración escolar. El progreso continuo en tecnología promete una adopción más

ISSN: 2810 – 8493
https://revista.ectperu.org.pe/index.php/ect/index

amplia y una mejora en las interacciones. El primer chatbot, ELIZA, fue creado por Joseph Weizenbaum en 1966 y utilizaba patrones predefinidos para simular una conversación terapéutica. A pesar de sus limitaciones, ELIZA demostró el potencial de la interacción hombre-máquina. En los años siguientes, el avance de la inteligencia artificial y el procesamiento del lenguaje natural impulsó la evolución de los chatbots. Programas como PARRY, desarrollado en 1972, y ALICE, creado en 1995, introdujeron técnicas más sofisticadas para generar respuestas. Sin embargo, el verdadero punto de inflexión llegó con el desarrollo de modelos de aprendizaje profundo, como GPT-3 de OpenAI en 2020, que permitieron a los chatbots comprender y generar texto con una precisión y coherencia sin precedentes. Hoy en día, los chatbots se utilizan en una amplia variedad de aplicaciones, desde la atención al cliente hasta la educación y la asistencia sanitaria, demostrando ser herramientas valiosas para mejorar la eficiencia y accesibilidad en múltiples sectores.

[2] nos dice que GPT es un modelo de lenguaje desarrollado por OpenAI y ChatGPT es un chatbot basado en este modelo. Este último está adquiriendo una relevancia creciente como asistente en áreas como la salud y la educación. Sin embargo, es crucial entender cómo se puede implementar efectivamente, cómo se pueden evaluar sus resultados en diversas tareas y cuáles son las limitaciones o problemas que podrían surgir durante su implementación.

El objetivo de la presente investigación es realizar una revisión sistemática de la literatura referente al desarrollo e implementación de chatbots que usan el modelo de lenguaje GPT para dar una visión de los problemas que surgen al usar el modelo, conocer las métricas de validación empleadas para medir el desempeño de la herramienta y conocer las tecnologías relacionadas con la implementación de chatbots con GPT. Asimismo, se implementó un caso práctico de la elaboración de un chatbot para responder las preguntas frecuentes de una cadena de supermercado.

2. Metodología de la Revisión Sistemática

El presente trabajo sigue una metodología basada en la propuesta de [3], la cual define un proceso de revisión sistemática de la literatura para investigación en ingeniería de software. Esta metodología se divide en tres etapas:

- **Etapa 1 - Planificación de la revisión sistemática**: Empieza con la definición de las preguntas de investigación, a partir de las cuales se extraen palabras clave para realizar la búsqueda de artículos. Antes de ejecutar la búsqueda, se establecen criterios de inclusión y exclusión para tomar o descartar artículos.

- **Etapa 2 - Ejecución del desarrollo de la revisión sistemática**: Se ejecuta la búsqueda en los sitios definidos usando las palabras clave y cadenas de búsqueda. Se seleccionan los artículos que cumplen todos los criterios definidos en la etapa anterior.

- **Etapa 3 - Análisis de los resultados de la revisión sistemática**: Se extraen las respuestas a las preguntas de investigación de cada artículo recogido en la fase anterior. Se consideraron artículos que respondieron al menos una pregunta.

La metodología para este trabajo es descriptiva basada en la revisión sistemática de artículos científicos**.**

3. Desarrollo

3.1. Formulación de las preguntas de investigación

Para la revisión de la literatura, se formularon tres preguntas de investigación relacionadas con el GPT. Estas preguntas están claramente definidas en la tabla 1.

Tabla 1. Preguntas de investigación

ID	Preguntas de investigación
P1	¿Cuáles son las principales tecnologías relacionadas con la implementación de chatbots con GPT?

ISSN: 2810 – 8493
https://revista.ectperu.org.pe/index.php/ect/index

P2	¿Cuáles son los principales problemas o limitaciones asociados a la implementación de chatbots con GPT?
P3	¿Cuáles son los criterios o métricas para evaluar el rendimiento de un chatbot con GPT?

3.2. Bases de datos y cadenas de búsqueda utilizadas

Tabla 2. Repositorios y cadenas de búsqueda usados

Repositorio	**Cadena de Búsqueda**
Scopus	**(chatbot OR "conversational agent" OR "conversational AI") AND (GPT OR ChatGPT)**
Web of Science	**(TI=(chatgpt) OR TI=(chatbot)) AND TI=(shop)**
Reaxys	**chatbot AND gpt AND implementation**
Google Scholar	**"gpt" and "chatbot" and "fine-tune"**
	"gpt" and "chatbot" and "interface"
ScienceDirect	**"gpt" and "chatbot" and "assistant" and "implementation"**
Wiley	**"gpt" and "chatbot" and "assistant"**
IEEE Access	**“gpt” AND (“chatbot” OR “conversational agent”) AND “implementation”**

3.3. Criterios de inclusión

Tabla 3. Criterios de inclusión para las investigaciones

Identificador	**Criterio de Inclusión**
CI-1	Los trabajos de investigación fueron publicados entre 2020 y 2024
CI-2	Los trabajos de investigación están escritos en Inglés o Español
CI-3	Los trabajos de investigación están disponibles o se tiene acceso a través de la institución
CI-4	Los trabajos de investigación tratan la implementación de chatbot

ISSN: 2810 – 8493
https://revista.ectperu.org.pe/index.php/ect/index

3.4. Resultados de la investigación

Tabla 4. Investigaciones recuperadas

ID	Título	Autor	Fecha de publicación
P1	A Generative Pretrained Transformer (GPT)–Powered Chatbot as a Simulated Patient to Practice History Taking: Prospective, Mixed Methods Study	Holderried, F., Stegemann–Philipps, C., Herschbach, L., Moldt, JA., Nevins, A., Griewatz, J., Holderried, M., Herrmann-Werner, A., Festl-Wietek, T., Mahling, M.	2024
P2	Developing a Medical Chatbot: Integrating Medical Knowledge into GPT for Healthcare Applications	Gams, M., Smerkol, M., Kocuvan, P., Zadobovšek, M.	2024
P3	Human Mimic Chatbot	Saradhi, M.V., Gaddampally, S., Chamarla, S., Cheluveru, A., y Tamarapu, A.	2023
P4	ReCo.ai: Using Generative Pre-trained Transformer 3 Model for a Chatbot in Answering Grade 10 Mathematics Questions.	S. Amado, P. T., T. Delos Santos, L. C., D. Germino, M. B., G. Nolos, A. G., C. Fronteras, V. C., C. Tria, R. L., T. Congreso, K. P., Oriol, D.	2023
P5	Making humanoid robots teaching assistants by using natural language processing (NLP) cloud-based services	Lekova, A., Tsvetkova, P., Tanev, T., Mitrouchev, P., y Kostova, S.	2022
P6	Revolutionizing customer interactions: Insights and challenges in deploying ChatGPT and generative chatbots for FAQs.	Khennouche, F., Elmira, Y., Djebaric, N., Himeurd, Y., & Amira, A.	2023
P7	A Chatbot for Collecting Psychoecological Data and Providing QA Capabilities.	Ni, Y., Chen, Y., Ding, R., & Ni, S.	2023
P8	Chatbot for assessing system security with OpenAI GPT-3.5	Lempinen, M., Pyyny, E., & Juntunen, A.	2023
P9	Data Augmentation and Preparation Process of PerInfEx: A Persian Chatbot With the Ability of Information Extraction	Safari, P., & Shamsfard, M.	2024
P10	Comparative Analysis of Generic and Fine-Tuned Large Language Models for Conversational Agent Systems. Robotics	Villa, L., Carneros-Prado, D., Dobrescu, C. C., Sánchez-Miguel, A., Cubero, G., & Hervás, R	2024
P11	Conversational Agent for Daily Living Assessment Coaching Demo	Finzel, R., Gaydhani, A., Dufresne, S., Gini, M., & Pakhomov, S.	2021
P12	Herding AI Cats: Lessons from Designing a Chatbot by Prompting	Zamfirescu-Pereira, J. D., Wei, H., Xiao, A., Gu, K., Jung, G., Lee, M. G.,	2023

	GPT-3	& Yang, Q.	
P13	The Adapter-Bot: All-In-One Controllable Conversational Model	Lin, Z., Madotto, A., Bang, Y., & Fung, P.	2020
P14	Audrey: A Personalized Open-Domain Conversational Bot	Hong, C. H., Liang, Y., Roy, S. S., Jain, A., Agarwal, V., Draves, R., Zhou, Z., Chen, W., Liu, Y., Miracky, M., Ge, L., Banovic, N. & Jurgens, D.	2020
P15	GastroBot: a chinese gastrointestinal disease chatbot based on the retrieval-augmented generation	Zhou, Q., Liu, C., Duan, Y., Sun, K., Li, Y., Kan, H., Gu, Z., Shu, J. & Hu, J.	2024
P16	Sentiment Analysis-Based Chatbot System to Enhance Customer Satisfaction in Technical Support Complaints Service for Telecommunications Companies	Juipa, A., Guzman, L., & Diaz, E.	2024
P17	The paradoxes of generative AI-enabled customer service: A guide for managers.	Ferraro, C., Demsar, V., Sands, S., Restrepo, M., & Campbell, C.	2024
P18	Beyond the Scalpel: Assessing ChatGPT's potential as an auxiliary intelligent virtual assistant in oral surgery.	Suárez, A., Jiménez, J., De Pedro, M. L., Andreu-Vázquez, C., García, V. D., Sánchez, M. G., & Freire, Y.	2024
P19	ChatGPT, the perfect virtual teaching assistant? Ideological bias in learner-chatbot interactions.	Van Poucke, M.	2024
P20	Development of an E-Commerce Chatbot for a University Shopping Mall.	Oguntosin, V., & Olomo, A.	2021
P21	Genuine2: An open domain chatbot based on generative models.	Rodríguez-Cantelar, M., de la Cal, D., Estecha, M., Gutiérrez, A. G., Martín, D., Milara, N. R. N., Martínez, R. & D'Haro, L. F.	2021

a. ¿Cuáles son las tecnologías relacionadas con la implementación de chatbots con GPT?

Aproximadamente el 24% de los estudios revisados implementan el modelo GPT-3.5 de OpenAI, destacándose principalmente la versión GPT-3.5-Turbo. Además, los modelos GPT se han implementado en una amplia variedad de formatos, incluyendo APIs, librerías y servicios en la nube. También se observan otros modelos derivados aparte de los de OpenAI, como DialoGPT de Microsoft, basado en GPT-2, que aunque menos utilizado, es reconocido en más de una investigación. En general, se pueden apreciar estas tecnologías en detalle aquí:

- Holderried et al. [5] implementaron el modelo GPT-3.5-Turbo de OpenAI, una variante de GPT-3.5, utilizando la Fetch API de JavaScript para realizar las solicitudes al modelo. Además, se desarrolló una interfaz mediante el uso de HTML y JavaScript, utilizando Tailwind CSS para la maquetación de los elementos.

- Gams et al. [6] implementaron el modelo GPT-4 de OpenAI a través de la librería LangChain. Además, se utilizó el modelo text-embedding-ada-002 de OpenAI para convertir textos relacionados con enfermedades en vectores. Estos vectores se almacenan en la base de datos Milvus para su posterior recuperación y uso.

ISSN: 2810 – 8493
https://revista.ectperu.org.pe/index.php/ect/index

- Saradhi et al. [7] implementaron el modelo DialoGPT de Microsoft en Python, utilizando librerías como Transformers, PyTorch y TensorFlow. Selenium también fue empleado específicamente para extraer mensajes de texto de WhatsApp.

- Amado et al. [8] implementaron el modelo GPT Text-Davinci-003 de OpenAI, una variante de GPT-3, utilizando JavaScript como lenguaje de programación.

- Lekova et al. [9] implementaron el modelo GPT-J-6B de EleutherAI en NLP Cloud. Además, se integraron funcionalidades como la conversión de voz a texto (Speech to Text) y de texto a voz (Text to Speech) utilizando IBM Watson. Esto permite transformar las voces de las personas a formato de texto y luego a audio, el cual es consumido por los robots NAO y EmoSan para que interactúen y respondan verbalmente. Todo este proceso está conectado a través de Node-RED.

- Ni et al. [10] implementaron modelos de lenguaje pre entrenados como GPT-3.5 basados en la arquitectura de Transformers.

- Lempinen et al. [11] implementó modelos de lenguaje avanzado basados, API de OpenAI, sistemas de detección de intrusiones basados en HOST y la tecnología de Prompts.

- Safari & Shamsfard [12] mencionan el uso de OpenAI's GPT-3.5-turbo para generar y parafrasear diálogos, el modelo Conditional BERT, que se utilizó para la clasificación de slots y la detección de intenciones, métodos de aumentos de datos y modelos de traducción automática para traducir grandes volúmenes de diálogos.

- Villa et al. [13] mencionaron el uso de Plataformas de Desarrollo de Chatbots (CDPs), Modelos de Lenguaje de Gran Escala (LLMs) y nodos de diálogo.

- Finzel et al. [14] implementaron GPT-2, dentro de la plataforma para agentes conversacionales MindMelt y servicios web de Python.

- Zamfirescu-Pereira et al. [15] implementaron un chatbot con GPT-3 con el enfoque de Investigación a través del Diseño y el uso extensivo de prompts.

- Lin et al. [16] implementaron DialoGPT, un modelo de agentes conversacionales basado en GPT-2 y usaron la plataforma BotUI para desplegar una versión demo del chatbot.

- Hong et al. [17] implementaron GPT-2 y el Amazon Conversational Bot Toolkit junto a funciones serverless de Amazon Web Services (AWS) en el marco del Amazon Alexa Prize Grand Challenge 3.

- Zhou et al. [18] implementaron GPT-3.5-turbo, el modelo de vectores de palabras GTE de Alibaba, la técnica de inteligencia artificial Retrieval-Augmented Generation, además de LlamaIndex y PDFReader.

- Juipa et al. [19] implementaron un chatbot con GPT-3.5 y WhatsApp Business como front-end para automatizar tareas de atención al cliente.

- Suárez et al. [21] implementaron Modelos de Lenguaje de Gran Escala (LLM), Interfaz de Programación de Aplicaciones (API) y técnicas de Procesamiento de Lenguaje Natural (PLN).

- Oguntosin & Olomo [23] mencionaron el uso de Python y React.js para desarrollar la interfaz y el backend del chatbot, la base de datos MySQL, también se emplean bibliotecas de Python como Spacy y APIs como Recast.ai para el entrenamiento y la comprensión del lenguaje natural y machine learning.

- Rodríguez-Cantelar et al. [24] implementaron GPT-2, DynamoDB y datasets de conversaciones en el marco del Alexa Socialbot Grand Challenge 4.

b. ¿Cuáles son los principales problemas o limitaciones asociados a la implementación de chatbots con GPT?

Uno de los problemas más mencionados es la falta o pérdida de contexto, posiblemente debido a conversaciones extensas o a entrenamientos insuficientes en ciertos temas. Esto podría llevar al modelo a proporcionar respuestas

ISSN: 2810 – 8493
https://revista.ectperu.org.pe/index.php/ect/index

incorrectas o vagas. Entre las limitaciones se encuentra el hecho de que estos modelos suelen requerir inversión para mejorar la latencia o la capacidad de retener contexto. En general, estos problemas y limitaciones pueden observarse en detalle aquí:

- En [5], se mencionó que debido a la estructura inicial de sus prompts, enfrentaron problemas con respuestas que se alejaban cada vez de la información de los pacientes, siendo que algunas respuestas estaban basadas en información ficticia, lo que afectaba la precisión y relevancia de los resultados proporcionados de la entrevista. También había problemas para que el chatbot se apegue a este rol de paciente.

- También [6] menciona que un desafío asociado con el uso de GPT-4 es que la calidad de las respuestas generadas puede verse afectada por la calidad de las preguntas realizadas. En particular, si las preguntas son imprecisas, ambiguas o poco detalladas, el modelo puede ofrecer respuestas de menor calidad o menos útiles.

- En [9], se mencionó que debido a que el servicio fue creado como una servicio cloud, enfrentaron problemas con una latencia de aproximadamente 100 ms. Esto se debió en parte al uso de una cuenta gratuita o de pago por uso. Además, al ajustar los parámetros como la temperatura del modelo, se notó que configurarla muy baja podría provocar la generación de bucles en los resultados. También se notó que el modelo podría generar contenido malicioso.

- Según [10], se tiene la necesidad de asegurar que los agentes conversacionales estén bien diseñados, efectivos y confiables, especialmente en el caso de agentes de detección de salud mental. Además, la interacción directa de los usuarios con modelos de lenguaje pre-entrenados puede generar respuestas que difieren significativamente de las expectativas del usuario.

- De acuerdo con [11], se presentaron obstáculos como la capacidad de memoria limitada ya que modelos como GPT-3.5 tienen una capacidad limitada para mantener contextos de conversación largos, también el afinado de modelos para tareas específicas y los costos y recursos computacionales.

- Se menciona en [12] la escasez de datos para entrenar el chatbot, ya que la lengua persa tiene poco recursos, problemas con el modelo generando respuestas genéricas como "sí" o "yo también" debido a la cantidad limitada de datos y diferencias en las fuentes de datos de NLU que afectan negativamente el rendimiento de la generación de lenguaje natural (NLG).

- Para [13], existe dependencia de datos de entrenamiento, que los modelos GPT pueden tener dificultades para comprender contextos específicos o dominios sin entrenamiento o ajuste fino específico y la gestión de entidades dinámicas.

- En [18], algunas de las respuestas del modelo implementado no son lo suficientemente específicas para resolver la duda enviada como prompt.

- El trabajo de [19] no menciona problemas de implementación, sin embargo se muestra que el chatbot propuesto no responde todas las dudas.

- En [20], se mencionan las relaciones comparadas de conectividad vs. aislamiento, costo reducido vs. precio alto, y finalmente, la alta calidad vs. falta de empatía.

- En [21], se observa que a pesar de su capacidad para simular conversaciones humanas, los chatbots GPT pueden generar respuestas convincentes pero incorrectas, conocidas como “alucinaciones”. Además,la implementación de chatbots GPT en la atención médica plantea preocupaciones éticas y legales relacionadas con la privacidad del paciente y la integridad de la práctica médica.

- Señala [22] entre los problemas y limitaciones encontrados: el sesgo ideológico, la falta de conciencia contextual, los desafíos éticos y los problemas de privacidad y datos..

- Finalmente, [23] enlista la comprensión textual, generación de respuestas, aprendizaje continuo y costos y recursos.

c. ¿Cuáles son los criterios o métricas para evaluar el rendimiento de un chatbot con GPT?

ISSN: 2810 – 8493
https://revista.ectperu.org.pe/index.php/ect/index

Generalmente, los criterios de evaluación incluyen la usabilidad, la consistencia y la precisión de las respuestas. Como métricas típicas se consideran un alto puntaje en cuestionarios administrados a los usuarios del chatbot, el tiempo de respuesta del chatbot, la precisión de las respuestas y otras medidas similares a las utilizadas en tareas de clasificación de machine learning. Es importante señalar que la evaluación de si una respuesta es aceptable para una pregunta puede requerir la participación de los autores o expertos. En general, estos criterios y métricas se pueden analizar en detalle aquí:

- En [5], se verificó la plausibilidad de las respuestas mediante una revisión realizada por dos autores de la investigación. Además, se utilizó un cuestionario de usabilidad para chatbots, adaptado de otra investigación, el cual se administró a los participantes del chatbot de entrevista. Finalmente, se analizó el porcentaje de preguntas que recibieron respuestas plausibles y no plausibles, así como el porcentaje de respuestas plausibles y no plausibles que estaban basadas en información ficticia o seguían la información dada del paciente.

- La propuesta de [6] fue evaluada mediante la intervención de los autores y expertos en el campo de la medicina verificando si las respuestas generadas eran adecuadas.

- En [7], la evaluación se realizó mediante una revisión detallada de las respuestas generadas por parte del chatbot frente a las preguntas de los contactos de WhatsApp, con el objetivo de determinar si el rendimiento del modelo era satisfactorio.

- En [8], se utilizó las métricas de evaluación siguientes: Accuracy siendo el número de respuestas correctas dividido por el número total de preguntas, Precision el número de respuestas correctas dividido por el número de respuestas totales. Recall el número de respuestas correctas dividido por el número de preguntas relevantes y F1-score combina precision y recall utilizando la fórmula F1 = 2 * (precision * recall) / (precision + recall) siendo en este contexto, precision = true positives/(true positives + false positives), y recall = true positives/(true positives + false negatives). También usando la estadística descriptiva (mean, standard deviation y frequency distribution) con las métricas anteriores para evaluar mejor los resultados.

- El trabajo de [10] incluye la evaluación de la usabilidad del sistema, la efectividad en la interacción con los usuarios, la satisfacción del usuario, la precisión en las respuestas generadas y la capacidad de mantener una conversación coherente y relevante.

- Por su parte, [11] evaluaron la precisión de respuestas, satisfacción del usuario, tiempo de respuesta y tasa de error.

- En [12] se consideró el F1-Score para clasificación de slots. El F1-Score es una medida de la precisión de un modelo de clasificación, especialmente útil cuando las clases están desequilibradas. Combina la precisión y la sensibilidad (recall) en una sola métrica mediante su media armónica. También tenemos la exactitud de coincidencia exacta y perplejidad.

- En [13] se evaluó la precisión de clasificación, exactitud de entidades y eficiencia operativa.

- Para [15], fue más importante evaluar los objetivos de UX (experiencia de usuario) de la investigación, usando tres ciclos iterativos de prompting para analizar las respuestas.

- En [16], se usaron las métricas Perplexity, BLEU, Avg-BLEU, n-gramas y entity-F1 para cada dataset de prueba.

- En [17], se realizó una encuesta en escala Likert (1-5) para medir la satisfacción de uso, además de tiempos de conversación promedio para distintos temas.

- Por su parte, [18] realizaron una evaluación humana de las respuestas generadas para medir la seguridad, la usabilidad y la fluidez del chatbot.

- En [19] se realizó una encuesta de escala Likert (1-5) con dos preguntas y se tomó el ratio de satisfacción de usuarios de acuerdo a dicha encuesta.

- En [20] se tienen presente la precisión de respuesta, comprensión contextual y tiempo de respuesta.

ISSN: 2810 – 8493
https://revista.ectperu.org.pe/index.php/ect/index

- En [21] se tienen presente la precisión, consistencia, comprensión y fiabilidad.
- Para [22], se presentan los siguientes criterios a evaluar: Precisión de respuestas, comprensión contextual, naturalidad del diálogo y tiempo de respuesta.
- En [23] se incluye la precisión y tiempo de respuesta.
- Finalmente, [24] evalúan la relación de las respuestas generadas con el tema de consulta, el grado de personalización de las respuestas y su calidad en términos de relevancia, naturalidad y compromiso

4. Caso Práctico

Nuestro equipo creó un Asistente Virtual para preguntas frecuentes para mejorar el proceso de atención al cliente en la plataforma web de la cadena de tiendas Supermercados Metro, usando el modelo GPT-3.5-Turbo elaborado por OpenAI. La elección de este modelo implicó contar con una cuenta de desarrollador en la plataforma de OpenAI y pagar $10 de crédito para el uso del modelo.

Se implementó utilizando una arquitectura Cliente/Servidor como muestra la Fig. 1. Para el servicio de hosting se optó por Firebase y PythonAnywhere. El fine-tuning del modelo GPT-3.5-TURBO se llevó a cabo utilizando Google Colab, empleando una hoja de cálculo para elaborar y almacenar las preguntas y respuestas, las mismas que fueron procesadas posteriormente para presentarse en un archivo JSONL como entrada para poder entrenar correctamente el modelo.

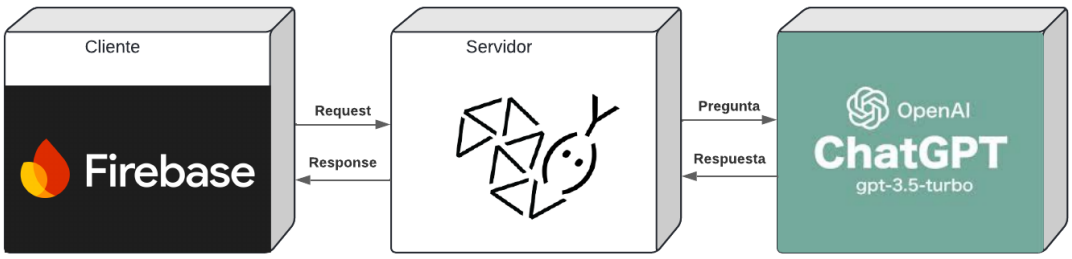

Figura 1. Arquitectura del Asistente Virtual

Empleamos la biblioteca openai junto con una “api_key” específica del proyecto generada en la plataforma OpenAI. Utilizando funciones de esta biblioteca, se cargó el archivo "Entrenamiento.jsonl" a OpenAI, especificando su uso para el

ISSN: 2810 – 8493
https://revista.ectperu.org.pe/index.php/ect/index

ajuste. Posteriormente, iniciamos el proceso de fine-tuning del modelo GPT-3.5-TURBO con el archivo cargado y esperamos la generación del nuevo modelo para su implementación.

Una vez obtenido el modelo, creamos el componente del servidor utilizando el framework Flask. Definimos una ruta "/chat" donde las peticiones tipo POST son dirigidas. Finalmente, para la parte del cliente, optamos por utilizar el framework de desarrollo Angular. Primero, se cargó la interfaz al repositorio del proyecto en GitHub. Luego, realizamos un despliegue inicial, lo que generó un archivo YAML que define el proceso automatizado de despliegue mediante GitHub Actions cada vez que se realiza un commit en el repositorio. Posteriormente, mediante StackBlitz creamos un servicio de envío de solicitudes POST al servidor en PythonAnywhere, encargado de recibir la respuesta para actualizar la interfaz. Para el diseño de la interfaz, adaptamos uno encontrado en línea [4]. La Figura 2 muestra el resultado final del Asistente Virtual.

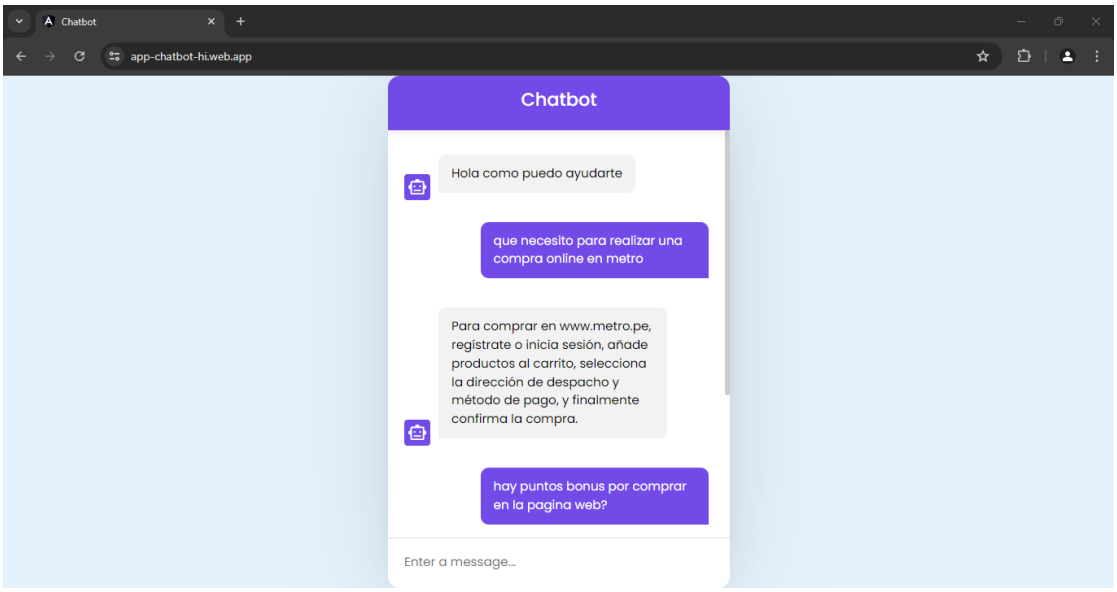

Figura 2. Ejemplo de uso del Asistente Virtual

5. Conclusiones

Desde el lanzamiento de las primeras versiones de GPT, la implementación de chatbots que usan esta familia de modelos ha sido continua y ha ido aumentando con el paso del tiempo y la mejora de los mismos. Este artículo de revisión sistemática ha proporcionado una visión integral sobre el desarrollo e implementación de chatbots utilizando modelos de lenguaje GPT. A través del análisis exhaustivo de la literatura reciente, se han abordado tres preguntas de investigación clave que guían este estudio:

- Tecnologías relacionadas con la implementación de chatbots con GPT: Los estudios revisados destacan una variedad de tecnologías utilizadas en la implementación de chatbots con GPT. La diversidad de enfoques subraya la flexibilidad y adaptabilidad de estos sistemas para diferentes contextos y aplicaciones.

- Problemas y limitaciones asociadas a la implementación de chatbots con GPT: Se identificaron varios desafíos críticos, como la generación de respuestas incoherentes o irrelevantes, o la necesidad de afinar los modelos para tareas específicas. Estos problemas subrayan la importancia de un diseño cuidadoso y la evaluación continua de estos sistemas para garantizar su efectividad y seguridad.

- Criterios y métricas para evaluar el rendimiento de los chatbots con GPT: Los criterios de evaluación incluyen precisión en la respuesta, satisfacción del usuario, tiempo de respuesta y la capacidad de mantener una conversación coherente. Estas métricas son fundamentales para medir la efectividad y usabilidad de los chatbots, proporcionando insights valiosos para su mejora continua y optimización.

En conclusión, los chatbots basados en GPT representan una poderosa herramienta que ha transformado significativamente diversas industrias, desde la atención al cliente hasta la educación y la salud. Sin embargo, su implementación efectiva requiere abordar los desafíos técnicos y éticos inherentes, así como la aplicación rigurosa de métricas de evaluación para

garantizar su desempeño óptimo y beneficios sostenibles. Futuras investigaciones pueden centrarse en resolver estos problemas pendientes para impulsar aún más el potencial de los chatbots en mejorar la interacción humano-máquina en diferentes contextos.

6. Referencias bibliográficas

[1] Khennouche, F., Elmira, Y., Djebaric, N., Himeurd, Y., & Amira, A. (2023). Revolutionizing customer interactions: Insights and challenges in deploying ChatGPT and generative chatbots for FAQs. Expert Systems with Applications, 1, 6. https://doi.org/10.1016/j.eswa.2024.123224

[2] Cheng, S. W., Chang, C. W., Chang, W. J., Wang, H. W., Liang, C. S., Kishimoto, T., Chang, J.P. C., Kuo, J.S. y Su, K. P. The now and future of ChatGPT and GPT in psychiatry. Psychiatry Clin. Neurosci., 2023, 77(11), pp 592-596. https://doi.org/10.1111/pcn.13588

[3] B. Kitchenham y P. Brereton, «A systematic review of systematic review process research in software engineering», Information and Software Technology, vol. 55, n.o 12, pp. 2049-2075, dic. 2013, doi: 10.1016/j.infsof.2013.07.010.

[4] 'How to Create Working Chatbot in HTML CSS and JavaScript', https://www.codingnepalweb.com/create-chatbot-html-css-javascript/, accesado el 17 de junio de2014

[5] Holderried, F., Stegemann–Philipps, C., Herschbach, L., Moldt, JA., Nevins, A., Griewatz, J., Holderried, M., Herrmann-Werner, A., Festl-Wietek, T., Mahling, M. A Generative Pretrained Transformer (GPT)–Powered Chatbot as a Simulated Patient to Practice History Taking: Prospective, Mixed Methods Study. JMIR Med. Educ., 2024; 10. https://doi.org/10.2196/53961

[6] Gams, M., Smerkol, M., Kocuvan, P., Zadobovšek, M. Developing a Medical Chatbot: Integrating Medical Knowledge into GPT for Healthcare Applications. Intelligent Environments 2024: Combined Proceedings of Workshops and Demos & Videos Session, 2024, 33, pp. 88–97, http://dx.doi.org/10.3233/AISE240018.

[7] Saradhi, M.V., Gaddampally, S., Chamarla, S., Cheluveru, A., y Tamarapu, A. Human Mimic Chatbot. World Journal of Advanced Research and Reviews, 2023, 18. 1232-1239. https://doi.org/10.30574/wjarr.2023.18.3.1228.

[8] S. Amado, P. T., T. Delos Santos, L. C., D. Germino, M. B., G. Nolos, A. G., C. Fronteras, V. C., C. Tria, R. L., T. Congreso, K. P., Oriol, D.: ReCo.ai: Using Generative Pre-trained Transformer 3 Model for a Chatbot in Answering Grade 10 Mathematics Questions. De La Salle University Senior High School Research Congress, June 2023

[9] Lekova, A., Tsvetkova, P., Tanev, T., Mitrouchev, P., y Kostova, S. Making humanoid robots teaching assistants by using natural language processing (NLP) cloud-based services. Journal of Mechatronics and Artificial Intelligence in Engineering, 2022, 3(1), pp. 30–39, https://doi.org/10.21595/jmai.2022.22720.

[10] Ni, Y., Chen, Y., Ding, R., & Ni, S. (2023). Beatrice: A Chatbot for Collecting Psychoecological Data and Providing QA Capabilities. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive*

Environment (PETRA '23), July 05–07, 2023, Corfu, Greece. ACM, New York, NY, USA. https://doi.org/10.1145/3594806.3596580

[11] Lempinen, M., Pyyny, E., & Juntunen, A. (2023). Chatbot for assessing system security with OpenAI GPT-3.5. University of Oulu, Degree Programme in Computer Science and Engineering.

[12] Safari, P., & Shamsfard, M. (2024). Data Augmentation and Preparation Process of PerInfEx: A Persian Chatbot With the Ability of Information Extraction. IEEE Access, 12, 19158-19180. https://doi.org/10.1109/ACCESS.2024.3360863

[13] Villa, L., Carneros-Prado, D., Dobrescu, C. C., Sánchez-Miguel, A., Cubero, G., & Hervás, R. (2024). Comparative Analysis of Generic and Fine-Tuned Large Language Models for Conversational Agent Systems. Robotics, 13(5), 68. https://doi.org/10.3390/robotics13050068

[14] Finzel, R., Gaydhani, A., Dufresne, S., Gini, M., & Pakhomov, S. (2021, April). Conversational agent for daily living assessment coaching demo. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 321-328). https://doi.org/10.18653/v1/2021.eacl-demos.38

[15] Zamfirescu-Pereira, J. D., Wei, H., Xiao, A., Gu, K., Jung, G., Lee, M. G., & Yang, Q. (2023, July). Herding AI cats: Lessons from designing a chatbot by prompting GPT-3. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (pp. 2206-2220). https://doi.org/10.1145/3563657.3596138

[16] Lin, Z., Madotto, A., Bang, Y., & Fung, P. (2021, May). The adapter-bot: All-in-one controllable conversational model. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 18, pp. 16081-16083). https://doi.org/10.1609/aaai.v35i18.18018

[17] Hong, C. H., Liang, Y., Roy, S. S., Jain, A., Agarwal, V., Draves, R., & Jurgens, D. (2020). Audrey: A personalized open-domain conversational bot. arXiv preprint arXiv:2011.05910. https://doi.org/10.48550/arXiv.2011.05910

[18] Zhou, Q., Liu, C., Duan, Y., Sun, K., Li, Y., Kan, H., ... & Hu, J. (2024). GastroBot: a Chinese gastrointestinal disease chatbot based on the retrieval-augmented generation. Frontiers in Medicine, 11, 1392555. https://doi.org/10.3389/fmed.2024.1392555

[11] Lempinen, M., Pyyny, E., & Juntunen, A. (2023). Chatbot for assessing system security with OpenAI GPT-3.5. University of Oulu, Degree Programme in Computer Science and Engineering.

[12] Safari, P., & Shamsfard, M. (2024). Data Augmentation and Preparation Process of PerInfEx: A Persian Chatbot With the Ability of Information Extraction. IEEE Access, 12, 19158-19180. https://doi.org/10.1109/ACCESS.2024.3360863

[13] Villa, L., Carneros-Prado, D., Dobrescu, C. C., Sánchez-Miguel, A., Cubero, G., & Hervás, R. (2024). Comparative Analysis of Generic and Fine-Tuned Large Language Models for Conversational Agent Systems. Robotics, 13(5), 68. https://doi.org/10.3390/robotics13050068

[14] Finzel, R., Gaydhani, A., Dufresne, S., Gini, M., & Pakhomov, S. (2021, April). Conversational agent for daily living assessment coaching demo. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 321-328). https://doi.org/10.18653/v1/2021.eacl-demos.38

[15] Zamfirescu-Pereira, J. D., Wei, H., Xiao, A., Gu, K., Jung, G., Lee, M. G., & Yang, Q. (2023, July). Herding AI cats: Lessons from designing a chatbot by prompting GPT-3. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (pp. 2206-2220). https://doi.org/10.1145/3563657.3596138

[16] Lin, Z., Madotto, A., Bang, Y., & Fung, P. (2021, May). The adapter-bot: All-in-one controllable conversational model. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 18, pp. 16081-16083). https://doi.org/10.1609/aaai.v35i18.18018

[17] Hong, C. H., Liang, Y., Roy, S. S., Jain, A., Agarwal, V., Draves, R., & Jurgens, D. (2020). Audrey: A personalized open-domain conversational bot. arXiv preprint arXiv:2011.05910. https://doi.org/10.48550/arXiv.2011.05910

ISSN: 2810 – 8493
https://revista.ectperu.org.pe/index.php/ect/index

[18] Zhou, Q., Liu, C., Duan, Y., Sun, K., Li, Y., Kan, H., ... & Hu, J. (2024). GastroBot: a Chinese gastrointestinal disease chatbot based on the retrieval-augmented generation. Frontiers in Medicine, 11, 1392555. https://doi.org/10.3389/fmed.2024.1392555

[19] Juipa, A., Guzman, L., & Diaz, E. (2024). Sentiment Analysis-Based Chatbot System to Enhance Customer Satisfaction in Technical Support Complaints Service for Telecommunications Companies. *ICSBT, 2024*, 28.

[20] Ferraro, C., Demsar, V., Sands, S., Restrepo, M., & Campbell, C. (2024b). The paradoxes of generative AI-enabled customer service: A guide for managers. Business Horizons. https://doi.org/10.1016/j.bushor.2024.04.013

[21] Suárez, A., Jiménez, J., De Pedro, M. L., Andreu-Vázquez, C., García, V. D., Sánchez, M. G., & Freire, Y. (2024). Beyond the Scalpel: Assessing ChatGPT's potential as an auxiliary intelligent virtual assistant in oral surgery. Computational And Structural Biotechnology Journal, 24, 46-52. https://doi.org/10.1016/j.csbj.2023.11.058

[22] Van Poucke, M. (2024). ChatGPT, the perfect virtual teaching assistant? Ideological bias in learner-chatbot interactions. Computers & Composition/Computers And Composition, 73, 102871. https://doi.org/10.1016/j.compcom.2024.102871

[23] Oguntosin, V., & Olomo, A. (2021). Development of an E-Commerce Chatbot for a University Shopping Mall. Applied Computational Intelligence And Soft Computing, 2021, 1-14. https://doi.org/10.1155/2021/6630326

[24] Rodríguez-Cantelar, M., de la Cal, D., Estecha, M., Gutiérrez, A. G., Martín, D., Milara, N. R. N., ... & D'Haro, L. F. (2021). Genuine2: An open domain chatbot based on generative models. *Proceedings Alexa Socialbot Grand Challenge SGC4*.