**Informe de Práctica Laboral**

**Estudiante: Claudia Queipo García**

**Tutor: Dr.C. Dionis López Ramos**

# Introducción

Durante el transcurso de mi inmersión en la práctica laboral, he sido partícipe en el desarrollo y perfeccionamiento del Sistema de Generación de Asistentes Virtuales. El presente informe tiene como propósito proporcionar una exposición exhaustiva de mi implicación en dicho proyecto, abordando aspectos que han sido determinantes en su evolución.

Iniciaré mi exposición explorando las Historias de Usuario del Sistema de Generación de Asistentes Virtuales, delineando con precisión las funcionalidades preceptivas que buscan optimizar la interacción entre los usuarios y los asistentes virtuales. Adicionalmente, se presentará una Propuesta de Interfaces detallada, subrayando la relevancia de un diseño que sea tanto intuitivo como eficiente en la experiencia del usuario.

A lo largo del corpus del informe, se incorporarán Diagramas de Secuencia y Actividades que desglosarán sistemáticamente cada uno de los flujos del sistema, proporcionando una aprehensión visual y lógica de su operatividad. Asimismo, se examinará el Diagrama de Estructura de Generación de Preguntas y Respuestas con Llama, profundizando en los matices técnicos que fundamentan la generación de contenido esencial.

La inclusión de una Propuesta de ChitChat destinada a todos los asistentes acentuará la amplitud de nuestro enfoque, considerando no solo la funcionalidad práctica, sino también la habilidad de los asistentes virtuales para participar en diálogos informales.

En un análisis crítico, se abordarán las obsolescencias inherentes a la tesis de Ernesto Duvalón, identificando las áreas en las cuales la investigación y el desarrollo del sistema han trascendido las expectativas iniciales.

Finalmente, se examinará la elección de Llama 2 sobre otros Modelos de Lenguaje de Aprendizaje (LLM), destacando las razones que respaldan esta determinación crucial y proporcionando una evaluación comparativa que fundamenta la superioridad de Llama 2 en el contexto de nuestro sistema de asistentes virtuales.

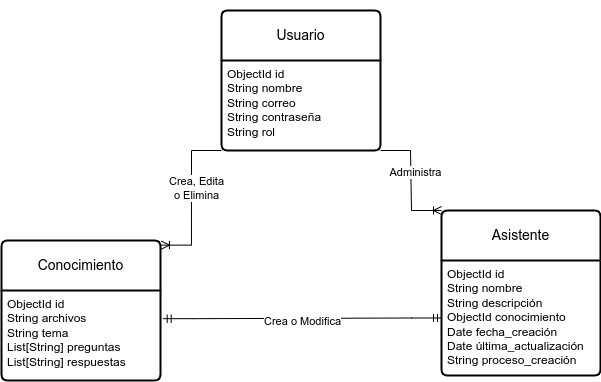
Este informe pretende constituir un análisis completo y minucioso de mi contribución al proyecto, procurando ofrecer una comprensión profunda de los elementos referidos y su aportación al desarrollo fructífero del Sistema de Generación de Asistentes Virtuales.

# Desarrollo

## Requisitos Funcionales

* RF 1: Registro del usuario usando su correo electrónico.
* RF 2: Inicio de sesión en el sistema, tanto con el correo electrónico como con la cuenta de Google o Facebook del usuario.
* RF 3: Generación de preguntas y respuestas a partir de un contexto dado.
* RF 4: Generar los archivos de conocimiento para el asistente virtual.
* RF 5: Entrenar al asistente.
* RF 6: Crear, editar, visualizar, eliminar y probar asistentes virtuales.

## Modelo Entidad Relación (MER)



## Actores del sistema

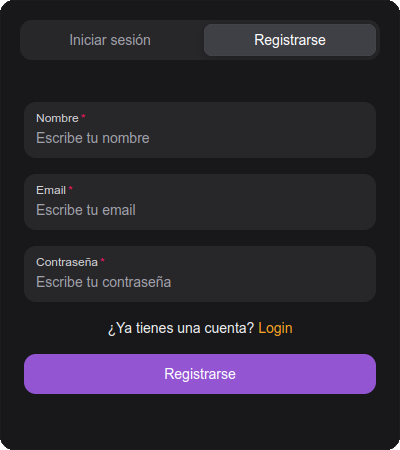
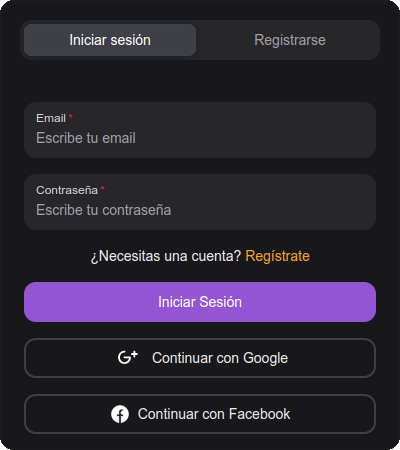
|  |  |
| --- | --- |
| Actor | Tareas que realiza |
| Cliente | Crea, edita, visualiza, elimina, prueba y entrena sus asistentes. |
| Administrador | Administra la base de datos del sistema, arregla fallos e implementa nuevas funcionalidades. |

## Propuesta de Interfaces.

En este diseño, más que en componentes visuales, se enfoca en la experiencia del usuario (UX, por sus siglas en inglés: User Experience). Por lo tanto, se realiza un maquetado (layout) de todas las posibles interfaces que pueda tener el asistente utilizando la herramienta Excalidraw. En el apartado de UI se propone el uso de la biblioteca de componentes visuales NextUI [1], la cual cuenta con una rica variedad de componentes visuales que permitirán acelerar el proceso de desarrollo de la interfaz, para así centrarme en el proceso de desarrollo de las funcionalidades.

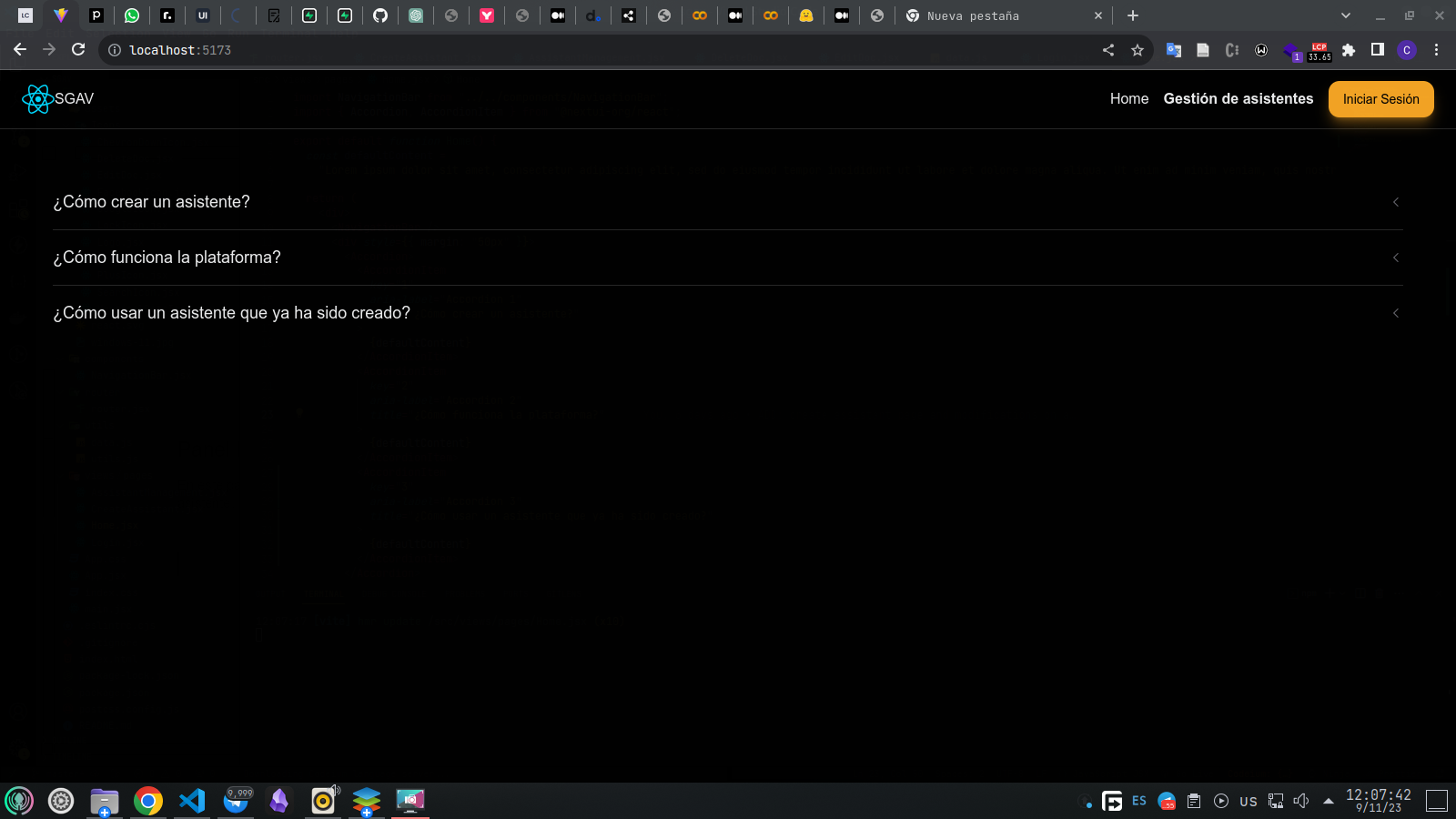
## Interfaces de inicio de Sesión y Registro de Usuario

Esta interfaz está basada en dos pestañas (tabs) de las cuales una pertenece al apartado de inicio de sesión y la otra al apartado de registro de usuario, este estilo optimiza el uso de nuevas rutas en la página a la vez que es una forma más intuitiva para el usuario.



## Panel Principal

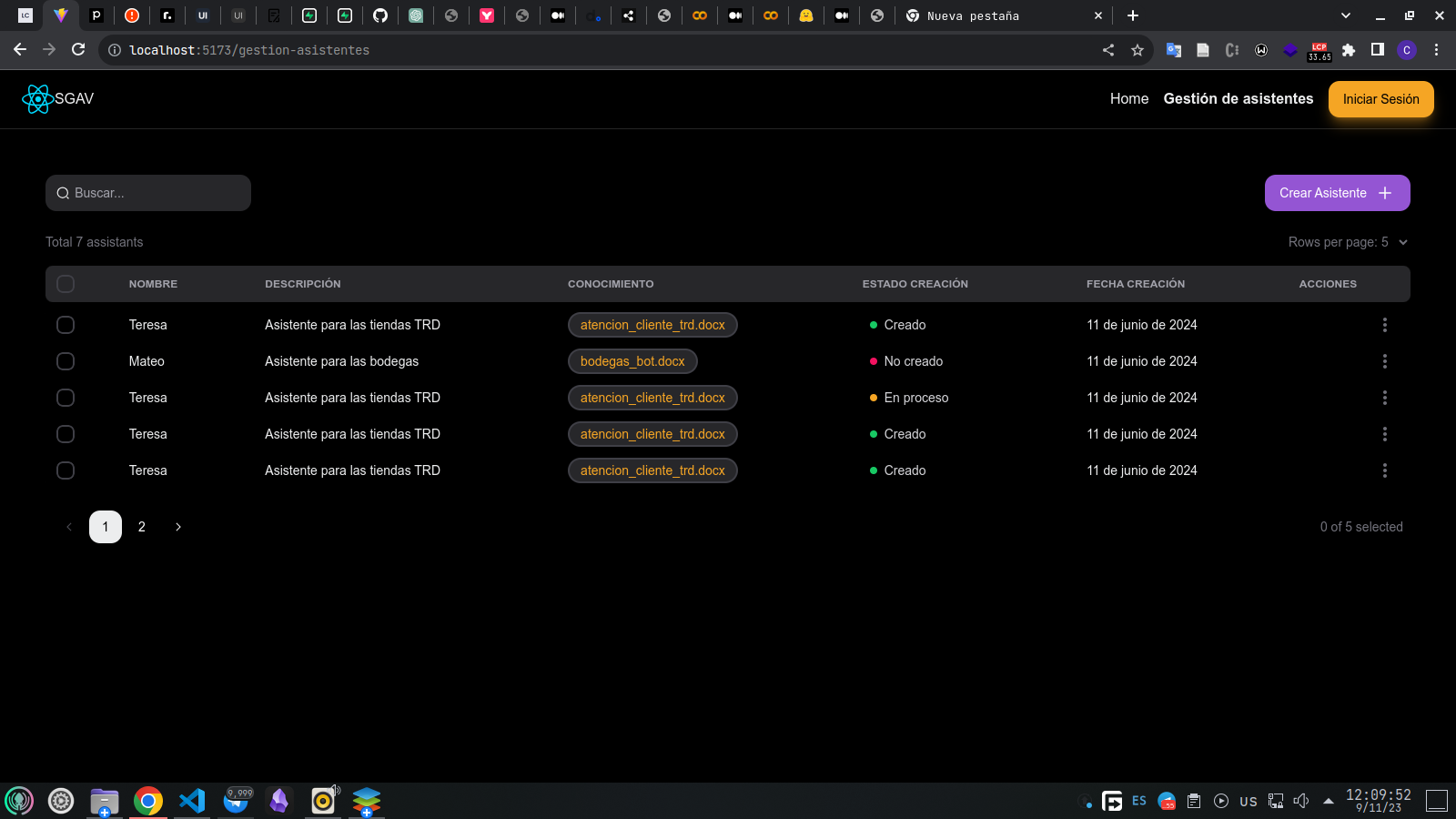
En este panel contará con una barra de navegación, una breve explicación de como funciona el asistente, a la derecha contará con una pequeña guía por pasos que indicará como crear correctamente un asistente.



## Interfaz del apartado de Administración de Asistentes

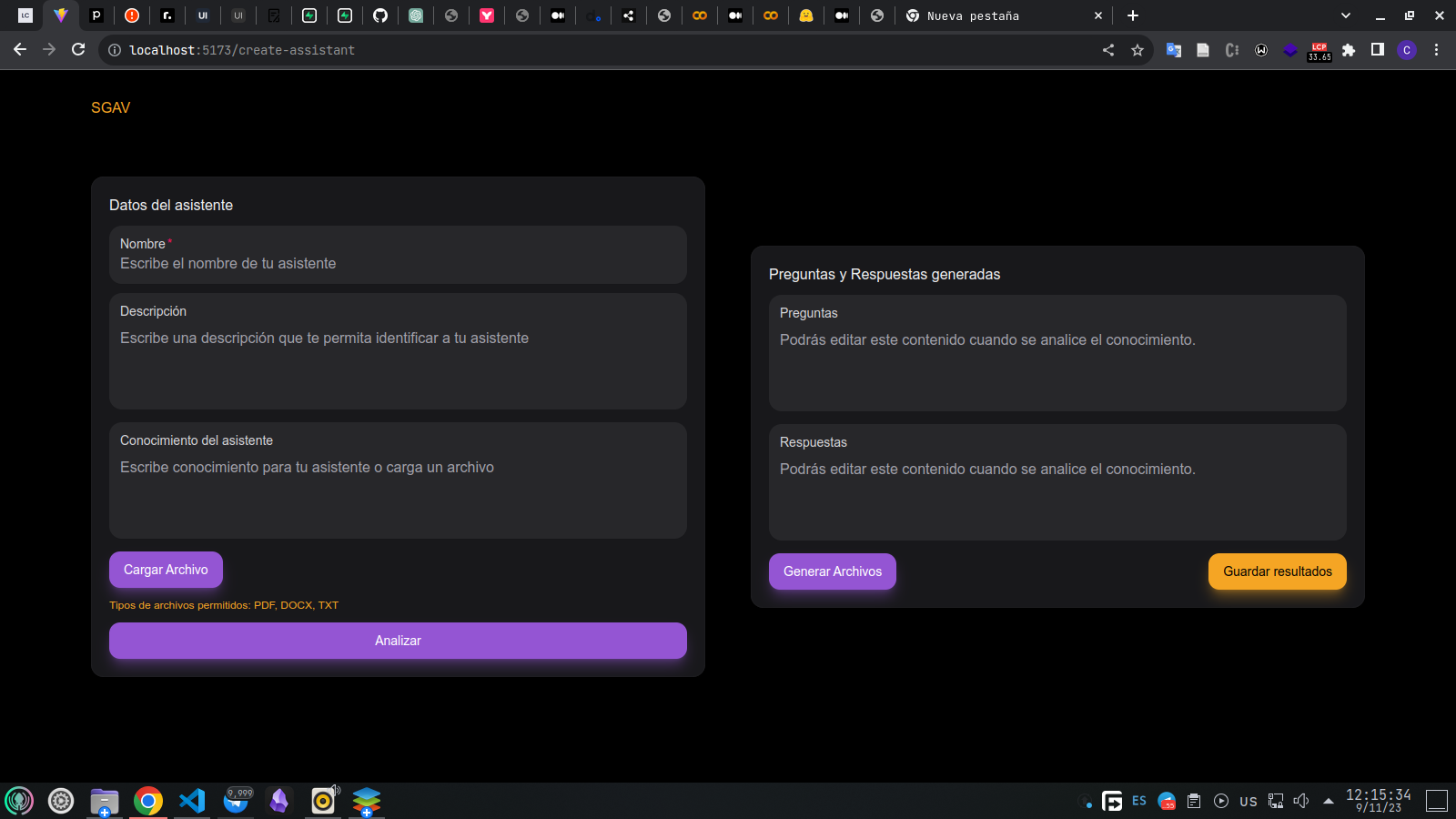
Esta interfaz contará con un botón para la creación del asistente, el cual al ser presionado abrirá un menú para la creación de un nuevo asistente, como se muestra en imágenes posteriores. También contará con un campo de búsqueda para poder filtrar los asistentes, ya sea por su nombre, fecha de creación, descripción, conocimiento o estado.

Cada asistente que sea creado mostrará en la tabla los campos que se observan y en el área de acción contará con un botón para realizar pruebas al asistente. A la derecha del mismo, habrá un botón para editar el asistente creado, el cual, al presionarlo, mostrará un menú de edición, tal como se muestra en imágenes posteriores. Por último, en el extremo derecho se encuentra el botón de eliminar, el cual elimina al asistente creado.



## Interfaces de creación y edición del asistente

Esta interfaz muestra dos formularios. El formulario de la izquierda es el que el usuario deberá rellenar, introduciendo el nombre, descripción y conocimiento del asistente. Estos dos últimos campos son opcionales. Por ejemplo, en lugar de introducir el conocimiento del asistente de forma manual, el usuario puede simplemente subir un archivo .pdf, .docx o .txt con el conocimiento del asistente. Luego, al presionar el botón "analizar", se generan preguntas y respuestas a partir de la información. Estas se colocarán en el formulario de la derecha, de manera que el usuario pueda editarlas en caso de que las generadas por el algoritmo no sean de su agrado. Después, el usuario puede generar los archivos de entrenamiento del asistente y/o guardar los resultados obtenidos en la generación de preguntas y respuestas.

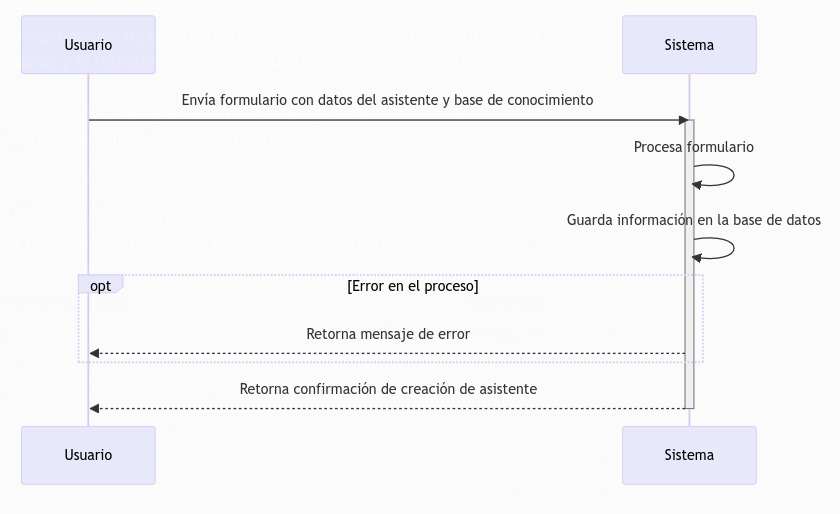


## Diagramas de Secuencia y Actividades de cada flujo del sistema

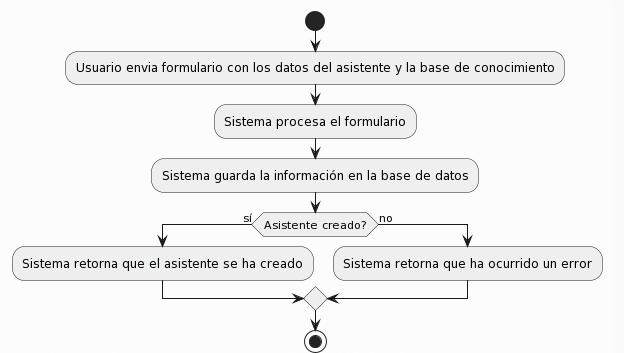
#### Creación de Asistente virtual

Como se observa en este flujo, primeramente, el usuario envía los datos del asistente que desea crear y la base de conocimiento que le va a otorgar. Luego, el sistema procesa el formulario, guarda la información en la base de datos. Si hubo un error en el proceso, notifica el error al usuario, y de no haber ninguno, confirma la creación del asistente.

##### Diagrama de Secuencia



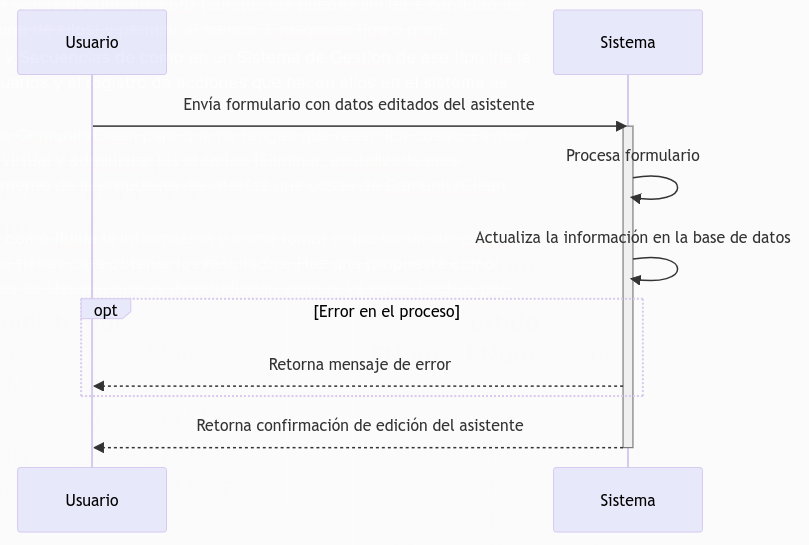
##### Diagrama de Actividades



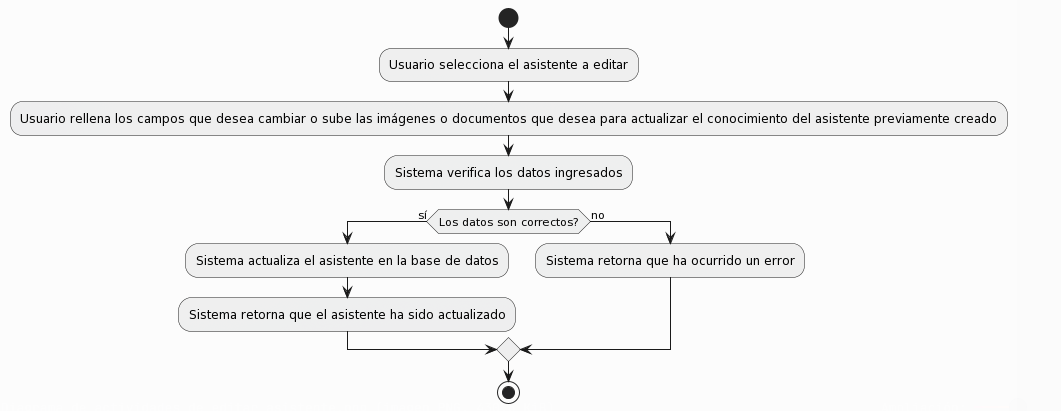
#### Edición de Asistente Virtual

El usuario primeramente envía el formulario con los datos que desea cambiar o el conocimiento que le quiere adicionar al asistente, este formulario es procesado, se actualiza en la base de datos el asistente, de ocurrir algún error en este proceso entonces se le notifica al usuario, de lo contrario se le notifica que no hubo ningún error retorna la confirmación de la edición correcta del asistente.

##### Diagrama de Secuencia



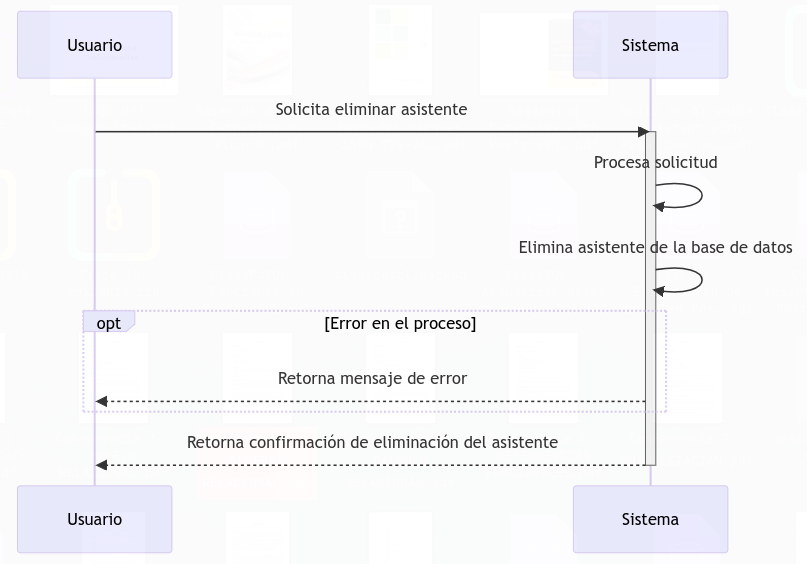
##### Diagrama de Actividades



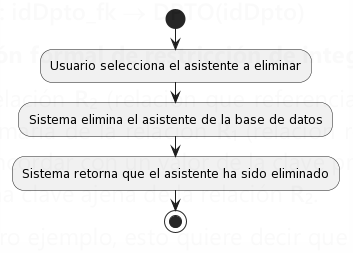
#### Eliminación de un asistente

El usuario realiza la solicitud de eliminación de su asistente, el sistema procesa la solicitud, luego se elimina el asistente de la base de datos, si ocurre algún error entonces este es notificado al usuario, de lo contrario se elimina el asistente y se le notifica al usuario.

##### Diagrama de Secuencia



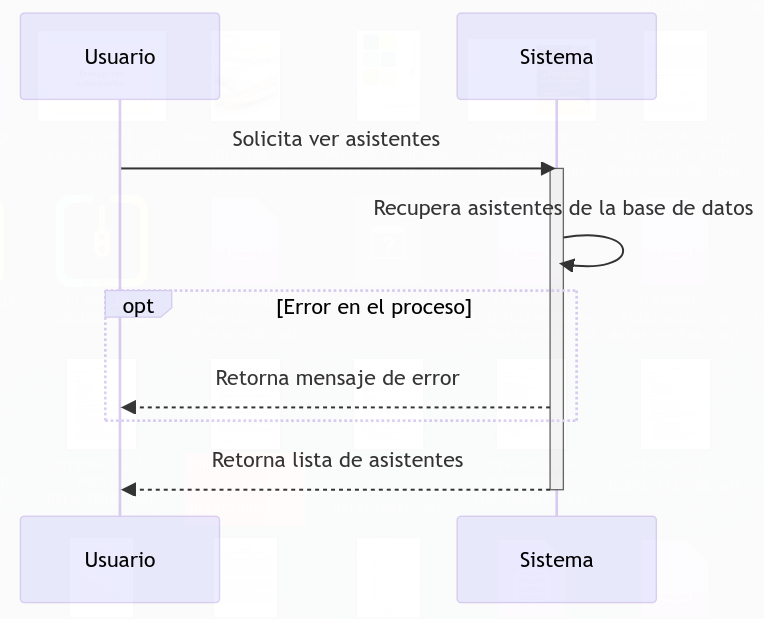
##### Diagrama de Actividades



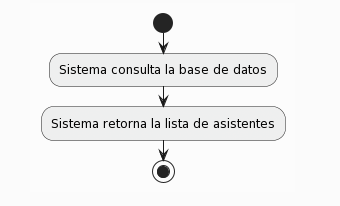
#### Visualizar Asistentes Creados

##### Diagrama de Secuencia

En este flujo el usuario presiona el apartado de Administrar Asistentes de el sistema, luego de esto el sistema recupera los asistentes creados y su información de la base de datos, de haber algún problema en la ejecución de esta demanda se arroja un mensaje de error, de lo contrario se le muestra al usuario en la tabla el listado de asistentes.



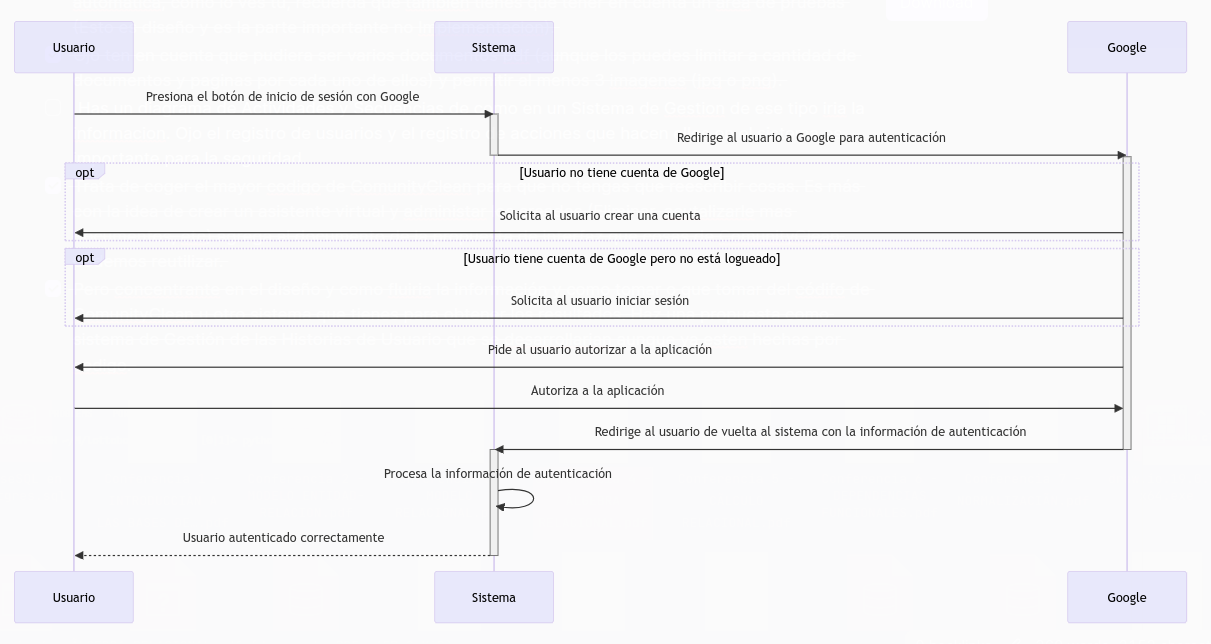
##### Diagrama de Actividades



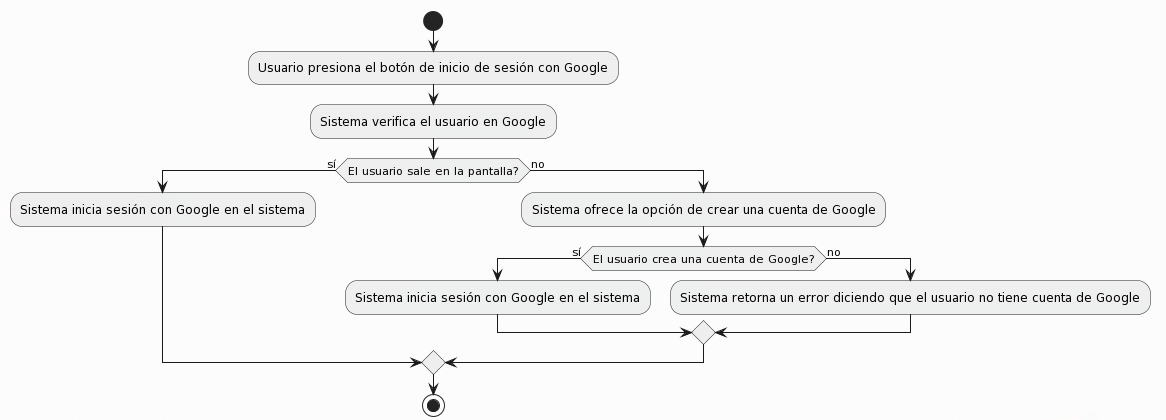
#### Inicio de Sesión con Google

La primera acción del usuario es presionar el botón de inicio de sesión con Google, luego este es redirigido a la vista de selección de cuentas de Google, si el usuario no posee una cuenta entonces debe solicitar crear una cuenta en Google, si ya posee una pero no encuentra una sesión activa con la misma entonces primeramente tiene que iniciar sesión con su cuenta, para luego autorizar su inicio de sesión en el sistema, si la autoriza entonces Google procesa la información de la autenticación, y luego se redirige el usuario a la pagina de el sistema y se le muestra una confirmación de su inicio de sesión satisfactorio.

##### Diagrama de Secuencia



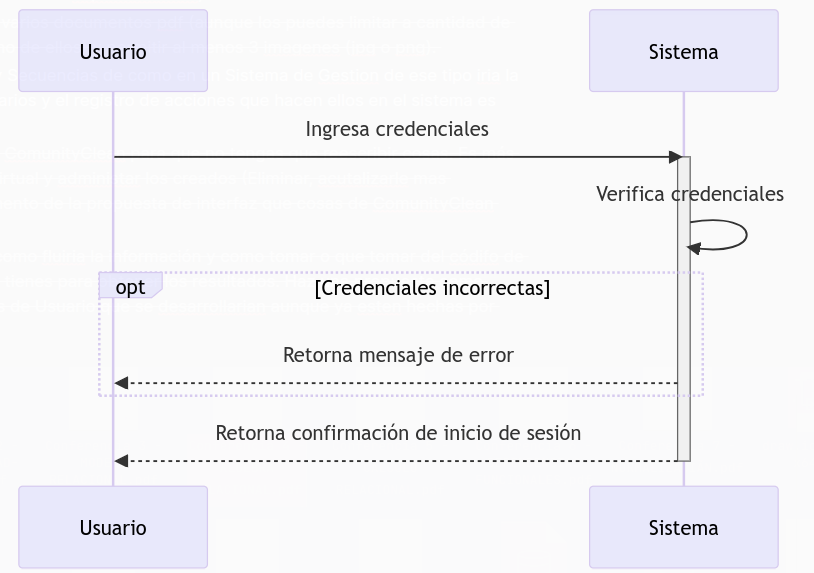
##### Diagrama de Actividades



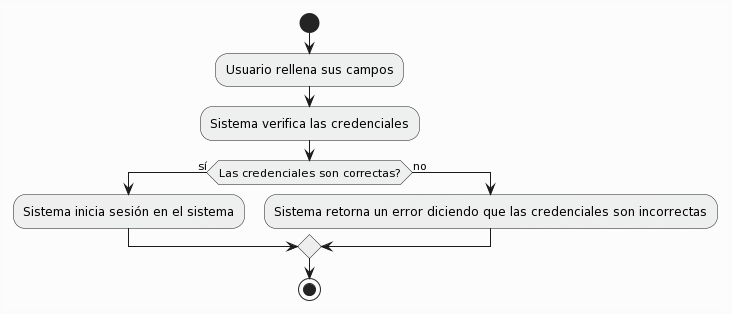
#### Inicio de sesión de Usuario

En el comienzo de este flujo el usuario ingresa sus credenciales, el sistema procede luego a verificar las credenciales ingresadas, de no coincidir con alguna de las que está en la base de datos, entonces envía la respuesta de credenciales incorrectas, de lo contrario retorna la confirmación del inicio de sesión satisfactorio.

##### Diagrama de Secuencia



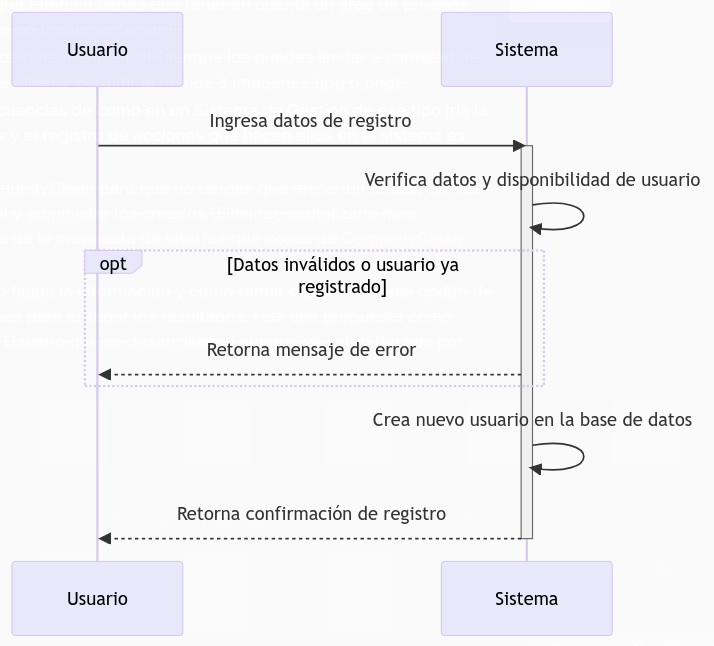
##### Diagrama de Actividades



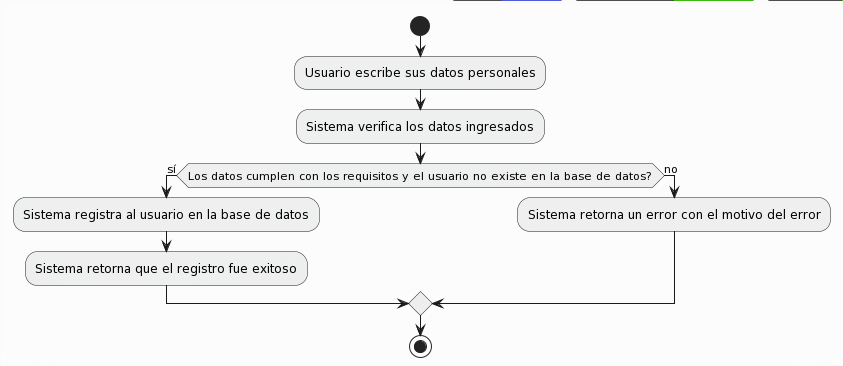
#### Registro de Usuario

Primeramente el usuario rellena el formulario de registro el cual le pide el correo, la contraseña y le pide que confirme la contraseña, luego se verifican los datos y la disponibilidad de ese usuario en la base de datos teniendo en cuenta como parámetro el correo proporcionado, si el usuario ya está registrado, emite un error y se lo notifica al usuario, de lo contrario se registra y luego envía la notificación de el registro exitoso.

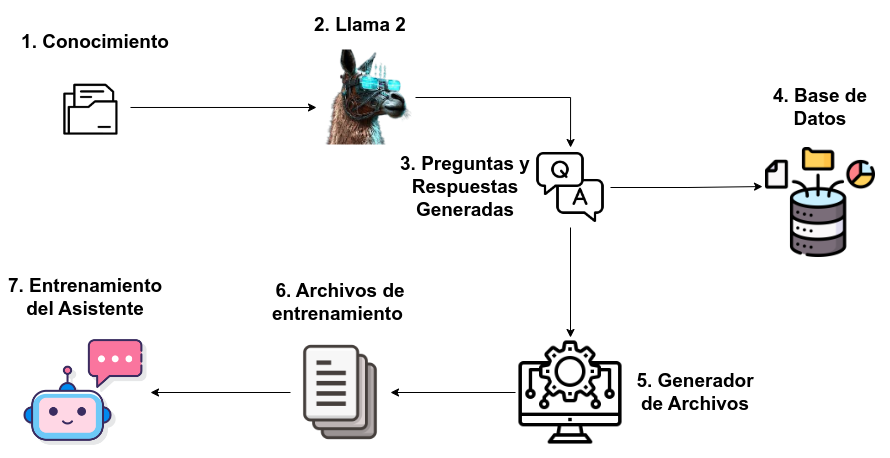
##### Diagrama de Secuencia



##### Diagrama de Actividades



## Diagrama de Estructura de Generación de Preguntas y Respuestas con Llama



## Propuesta de ChitChat para todos los asistentes

Esta propuesta sería básicamente una plantilla inicial con la cual contarían todos los asistentes que sean creados en la plataforma.

Esta sería la plantilla del asistente en lenguaje español.

Usuario: ¡Hola!

Asistente: ¡Hola! ¿En qué puedo ayudarte hoy?

Usuario: Adiós

Asistente: Adiós, ha sido un placer atenderte hoy.

**Saludos**:

1. ¡Hola! ¿En qué puedo ayudarte hoy?

2. Buenos días, ¿cómo puedo ayudarte?

3. ¡Hola! Es un placer estar aquí para asistirte.

4. ¡Hola! ¿En qué puedo colaborar contigo en este día?

5. ¡Saludos! ¿Cómo puedo ser de utilidad hoy?

**Despedidas**:

1. ¡Hasta luego! Si necesitas algo más, no dudes en preguntar.

2. Que tengas un excelente día. ¡Adiós!

3. ¡Hasta pronto! Estoy aquí si necesitas más ayuda en el futuro.

4. Ha sido un placer asistirte. ¡Adiós y que tengas un buen día!

5. ¡Adiós! Espero haber sido útil. Si tienes más consultas, estaré aquí para ayudarte.

Esta sería la plantilla en inglés.

User: Hello! Assistant: Hello! How can I assist you today? User: Goodbye Assistant: Goodbye, it has been a pleasure serving you today.

**Greetings**:

1. Hello! How can I assist you today?

2. Good morning, how may I assist you?

3. Hello! It's a pleasure to be here to assist you.

4. Hello! How can I collaborate with you today?

5. Greetings! How can I be of service today?

**Farewells**:

1. See you later! If you need anything else, do not hesitate to ask.

2. Have a great day. Goodbye!

3. See you soon! I am here if you need more help in the future.

4. It has been a pleasure assisting you. Goodbye and have a good day!

5. Goodbye! I hope I have been helpful. If you have more queries, I will be here to assist you.

## Análisis crítico de las obsolescencias de la tesis de Ernesto Duvalón

A lo largo de este año se ha realizado un gran avance en la inteligencia artificial, y como era de esperar el campo de el procesamiento del lenguaje natural, ha tenido un avance directamente proporcional al avance de la IA. En el caso de la forma en la que se generan las preguntas y respuestas, esta sería muy laboriosa, porque en este caso, se utilizarían dos modelos, un generador de preguntas y un generador de respuestas, otro punto a tener en cuenta es que estos modelos no tienen la capacidad de análisis que tienen los largos modelos del lenguaje como Llama 2 entonces las preguntas generadas siempre tendrán un formato predeterminado, por último la cantidad de parámetros con la que ha sido entrenado el modelo utilizado en esa tesis es de 248 Millones, mientras que el modelo de Llama 2 que se va a utilizar tiene 7 Billones de parámetros y este cuenta con el reconocimiento de varios idiomas, cosa de la cual el modelo de bert carece.

## Porqué Llama 2 sobre los otros LLM

Se podría pensar que la elección de Llama se basa en que es Open Source, pero no, la realidad es que Llama frente a modelos como GPT-3.5, PALM y FALCON, tiene un rendimiento bastante adecuado y las métricas que se muestran a continuación lo demuestran.

Si comparamos a Llama 2 con Falcon, como bien se observa en la siguiente foto, en el caso de los modelos de 7 Billones de parámetros Llama tiene casi el doble de puntuación en MMLU (Massive Multitask Language Understanding) métrica que se utiliza para medir cuán bien un modelo de lenguaje a gran escala (LLM) comprende el lenguaje y puede resolver problemas con el conocimiento adquirido durante su entrenamiento.

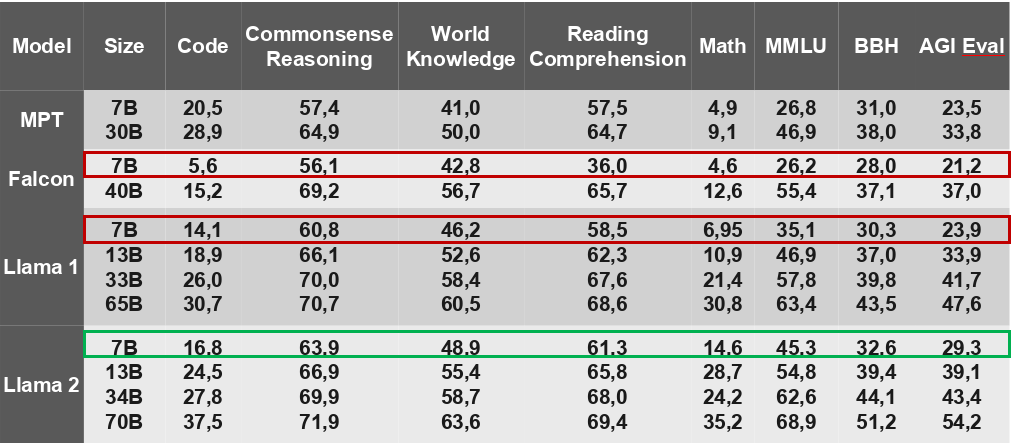


Tabla comparativa entre los modelos de Falcon y Llama [2]

En la próxima imagen, se observa como pese a que el modelo Llama 2 70 billones pese a ser el que menor puntuación tiene de sus contrincantes PALM 1, PALM 2, GPT-3.5, GPT-4, se muestra muy cercano en sus métricas a estos modelos.

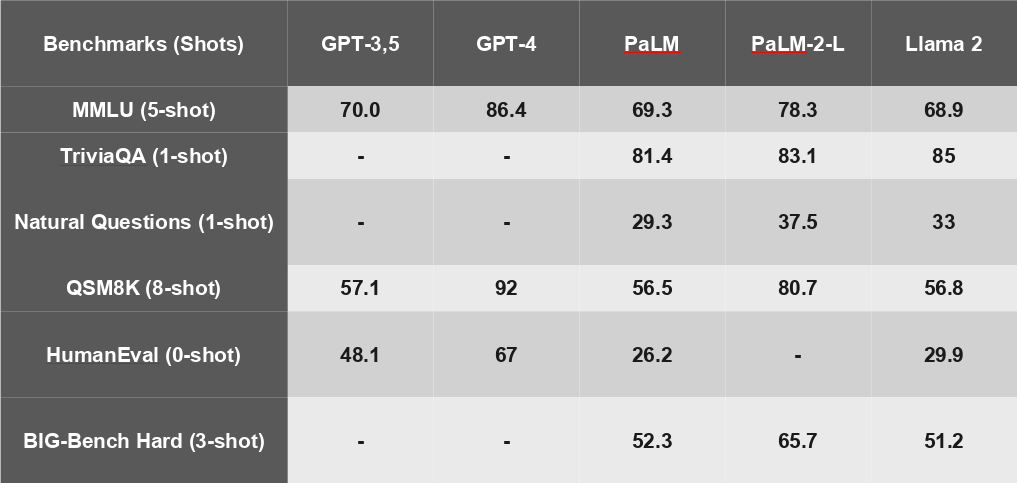


Tabla comparativa entre Llama2, PaLM, PaLM2, GPT-3.5 y GPT-4 [2]

Por último, el motivo de peso por el cual elijo a Llama 2 para el desarrollo de esta tesis es que se le puede realizar fine-tuning para lograr que el modelo realice el proceso de generación de preguntas y respuestas con una mayor precisión. Además, este proceso se podría llevar a cabo de forma local, ya que los requisitos de hardware de este modelo permiten ejecutarlo sin necesidad de una GPU. Estos dos aspectos son importantes, ya que los modelos anteriormente mencionados carecen de ellos, algunos modelos no permiten el fine-tuning por su licencia, mientras que otros son muy costosos en cuanto a recursos de hardware [3].

# Conclusiones

Con la culminación de este informe de práctica laboral, se llegan a las siguientes conclusiones:

1. Se definieron los requisitos funcionales del sistema.
2. Se definió el modelo entidad relación el cual servirá para tener una idea de los modelos de datos que tendrá el sistema y cuales de ellos deben persistir en una base de datos.
3. Se realizó el proceso de planificación y documentación del software mediante diagramas.
4. Se desarrollo el frontend del Sistema de Gestión de asistentes virtuales.
5. Se defninió el flujo que va a tener el algoritmo de generación de preguntas y respuestas con Llama 2.

# Referencias Bibliográficas

[1]jrgarciadev, “NextUI - Beautiful, fast and modern React UI Library,” NextUI. Accessed: Nov. 01, 2023. [Online]. Available: https://nextui.org  
   
 [2]S.-H. Tsang, “Brief Review — Llama 2: Open Foundation and Fine-Tuned Chat Models,” Medium. Accessed: Oct. 28, 2023. [Online]. Available: https://sh-tsang.medium.com/brief-review-llama-2-open-foundation-and-fine-tuned-chat-models-6666eb8b56b7  
   
 [3]S. Kazi, “Top Large Language Models (LLMs): GPT-4, LLaMA 2, Mistral 7B, ChatGPT, and More,” Vectara. Accessed: Nov. 02, 2023. [Online]. Available: https://vectara.com/top-large-language-models-llms-gpt-4-llama-gato-bloom-and-when-to-choose-one-over-the-other