**Informe Técnico**

**Estudiante: Claudia Queipo García**

**Fecha: 23-10-2023**

# Historias de Usuario

* Registrarse en el sistema usando su correo electrónico
* Iniciar sesión en el sistema tanto con email, como con su cuenta de google.
* Crear asistentes virtuales rellenando los campos de nombre, descripción, archivos de conocimiento e imágenes del asistente.
* Editar los asistentes creados.
* Visualizar los asistentes creados.
* Eliminar los asistentes creados
* Probar los asistentes

# Propuesta de Interfaces.

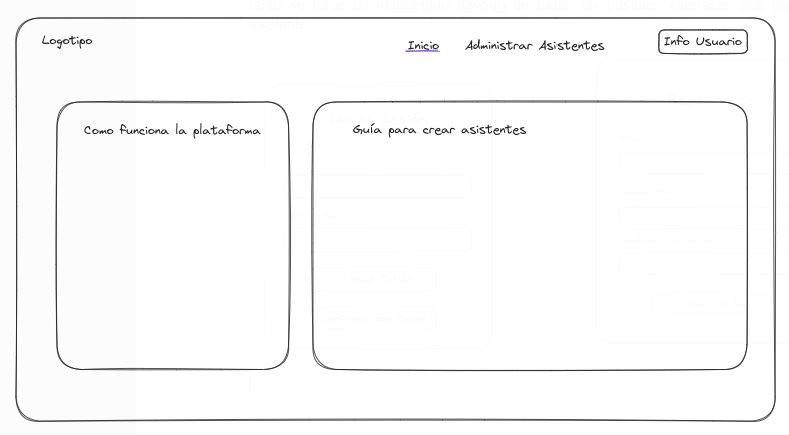
En este diseño más que en componentes visuales, se basa en la UX (User Experience), por lo tanto se hace un maquetado (layout) de todas las posibles interfaces que pueda tener el asistente usando la herramienta Excalidraw. En el apartado UI se propone el uso de la biblioteca de componentes visuales de ShadcnU [1]I la cual cuenta con una rica variedad de componentes visuales los cuales 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



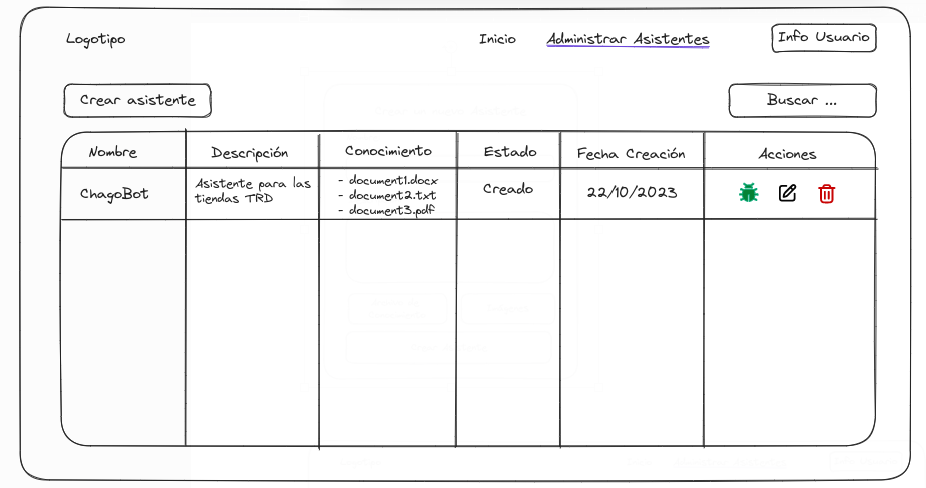
## Panel Principal

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.



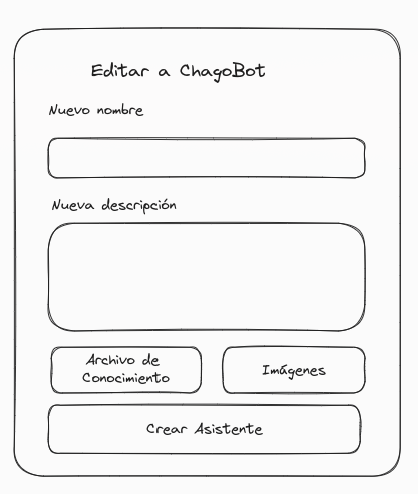
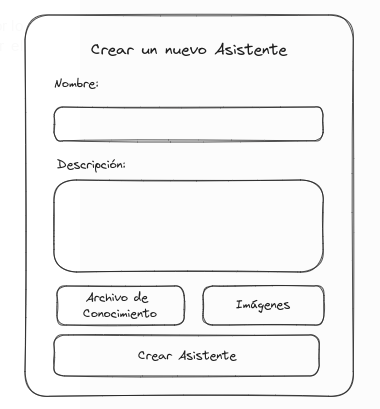
## Interfaz del apartado de Administración de Asistentes

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 bien 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 de el mismo 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

Estas interfaces muestran los campos nombre, descripción, archivos de conocimiento (Máximo 3 archivos de conocimiento para entrenar con ellos al asistente), imágenes (Máximo 3 imágenes ), por último contará con el botón de acción ya sea crear o editar



## Reutilización de Código:

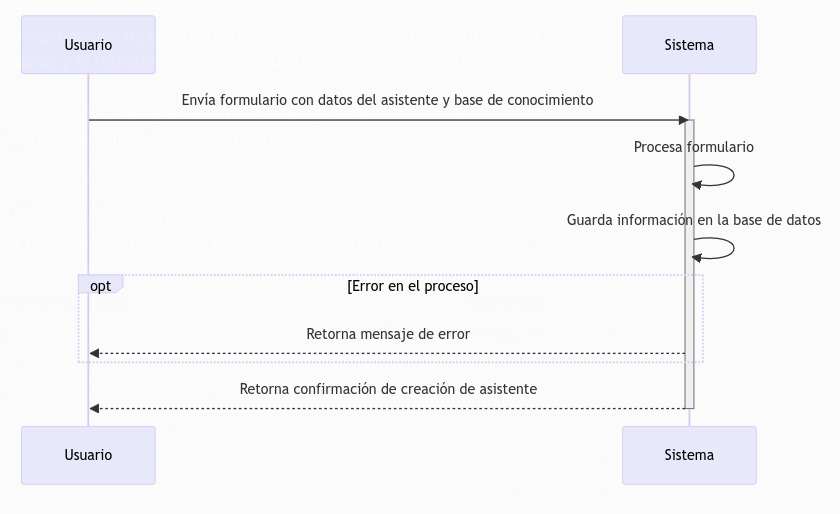
En este caso no voy a reutilizar código de ninguna plataforma dado que las interfaces tienen especificaciones muy particulares. En el caso de la plataforma City Clean no uso su codigo dado que en esta plantilla se utiliza MaterialUI la cual crea un bundle de Javascript demasiado grande para este proyecto que es más pequeño, otro de los motivos es que la plantilla de City Clean es extremadamente grande, lo que me reduciría tiempo de desarrollo tratando de entender alguno de sus componentes y como implementarlo.

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

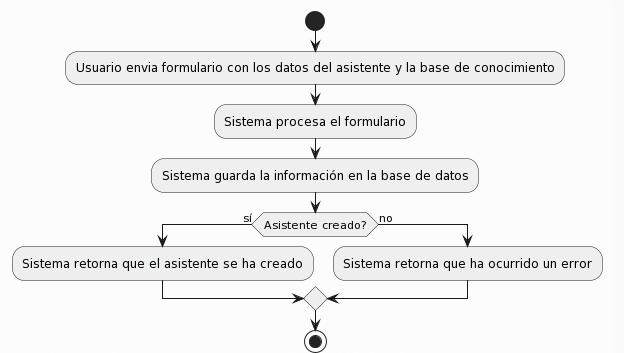
#### Creación de Asistente virtual

Como bien 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 entonces confirma la creación del asistente.

##### Diagrama de Secuencia



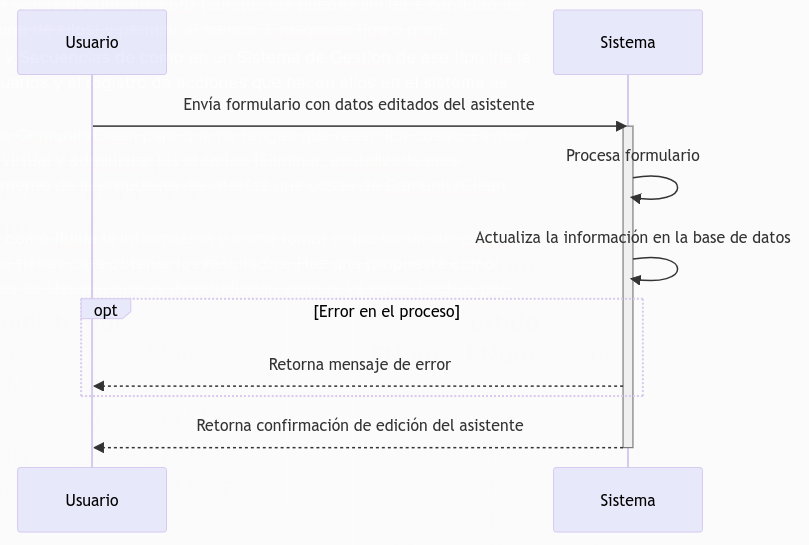
##### Diagrama de Actividades



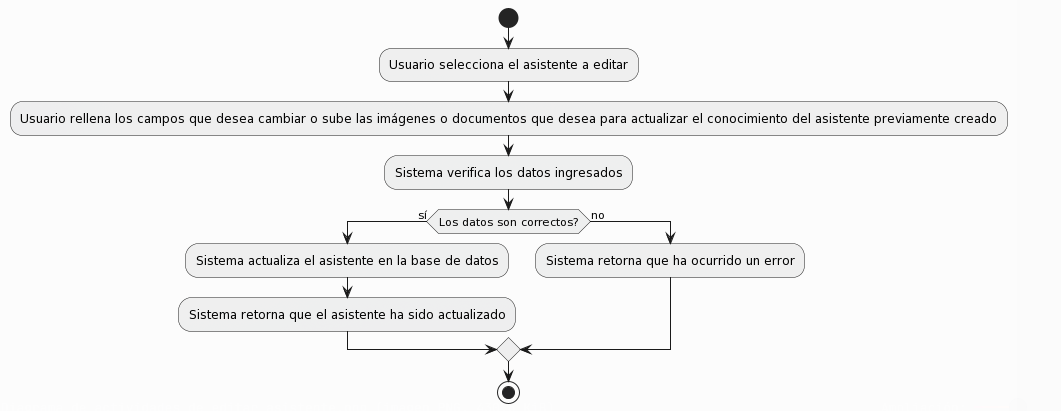
#### Edición de Asistente Virtual

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



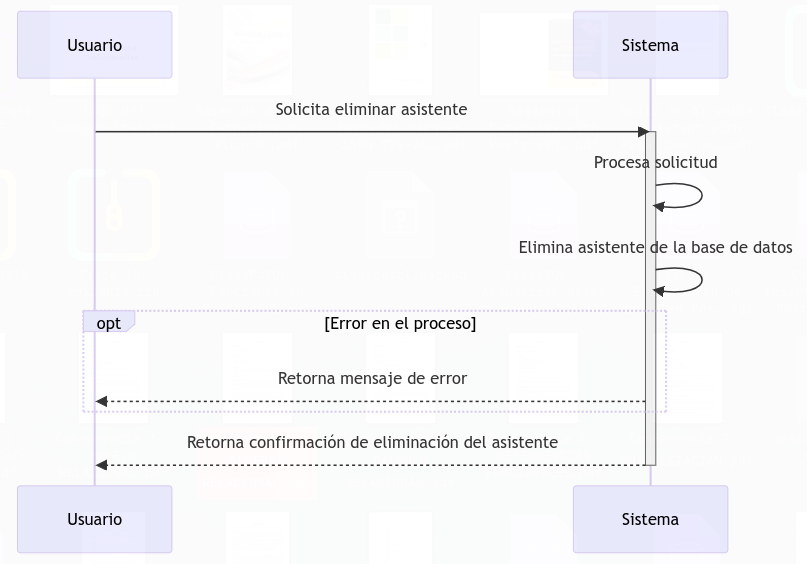
##### Diagrama de Actividades



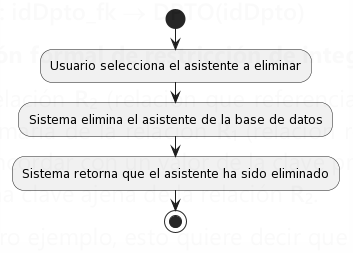
#### Eliminación de un asistente

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



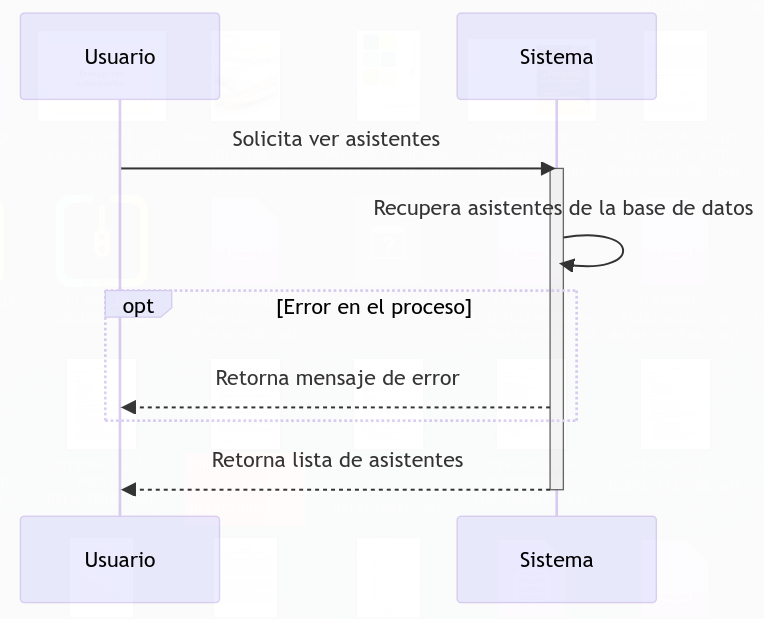
##### Diagrama de Actividades



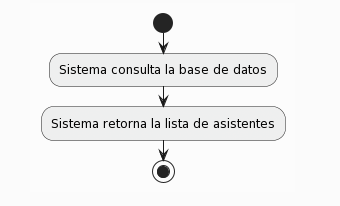
#### Visualizar Asistentes Creados

##### 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.



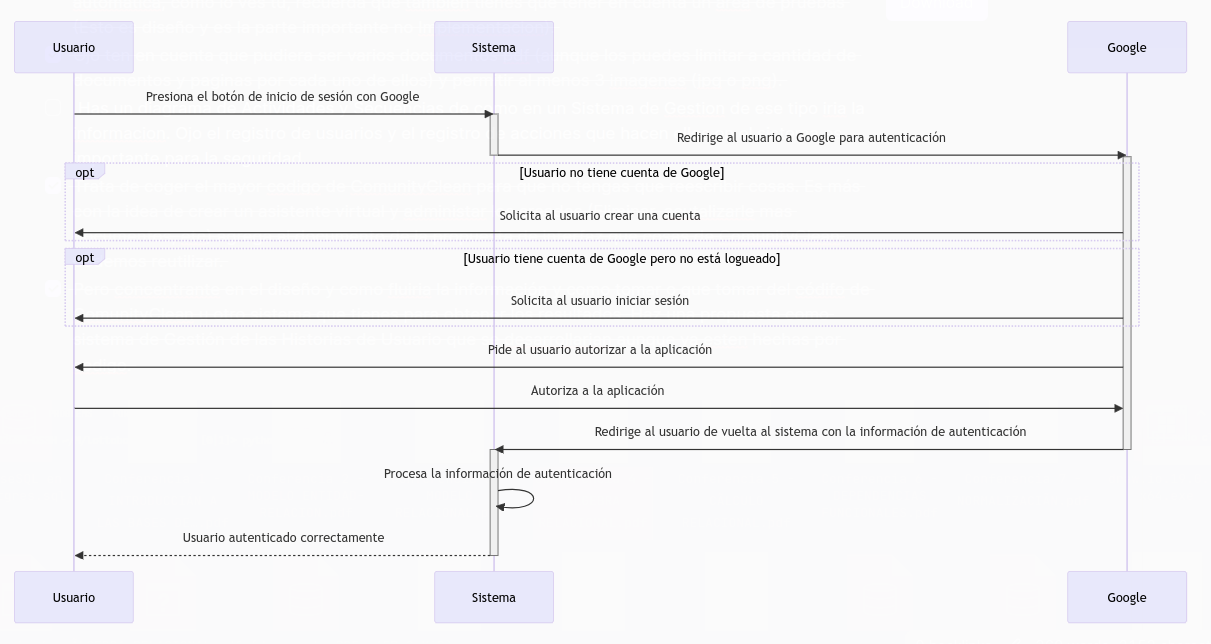
##### Diagrama de Actividades



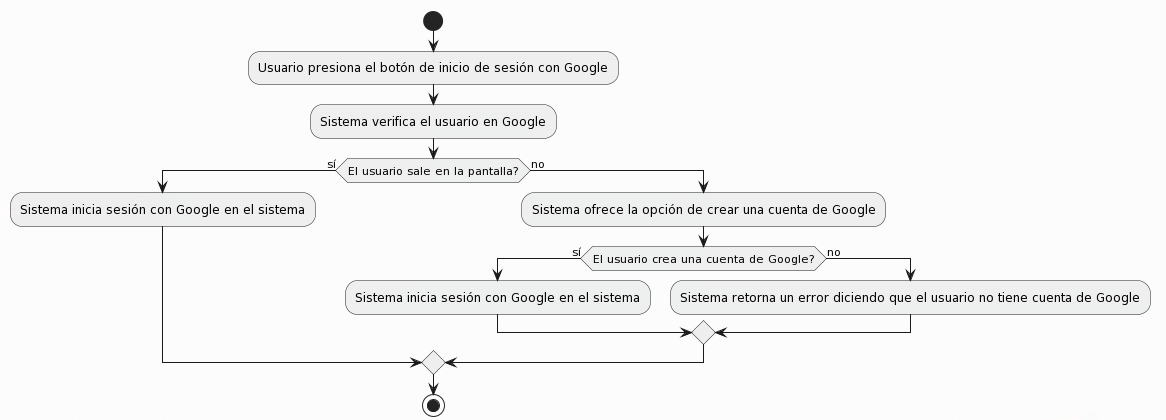
#### Inicio de Sesión con Google

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 se encuentra logeado 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



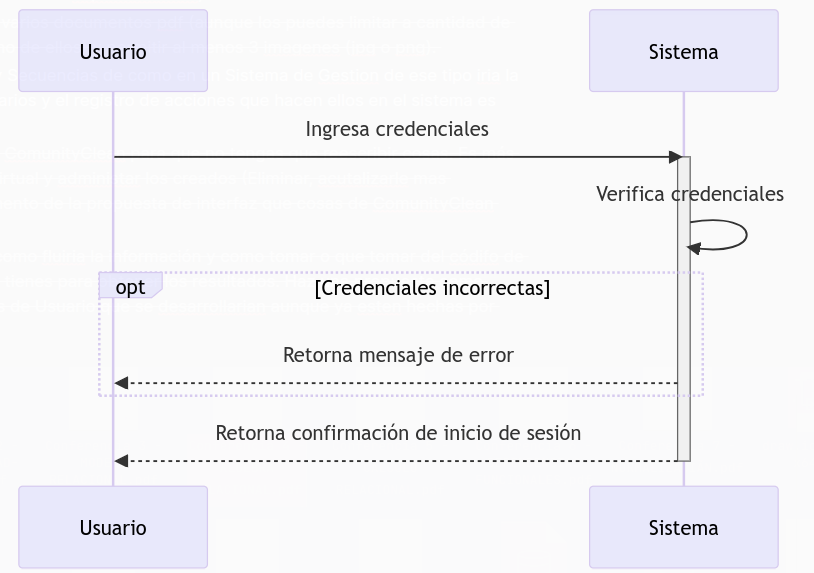
##### Diagrama de Actividades



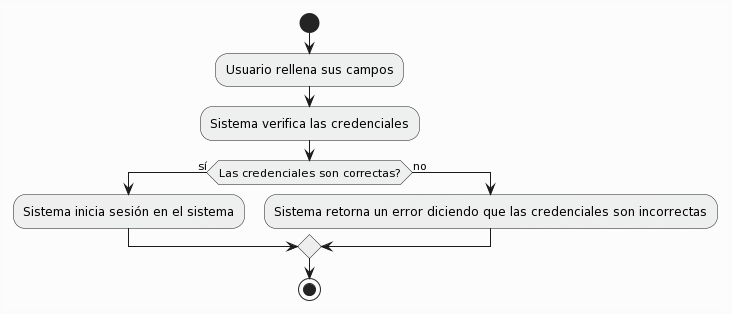
#### Inicio de sesión de Usuario

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



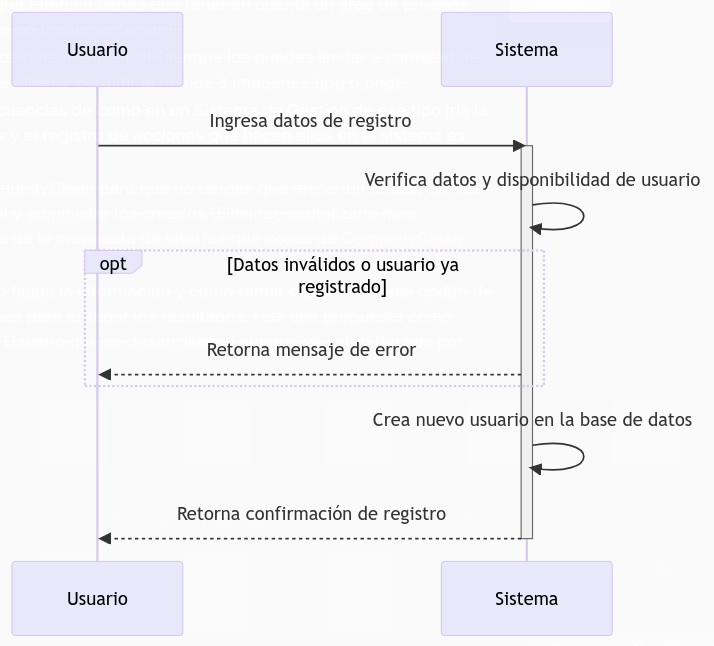
##### Diagrama de Actividades



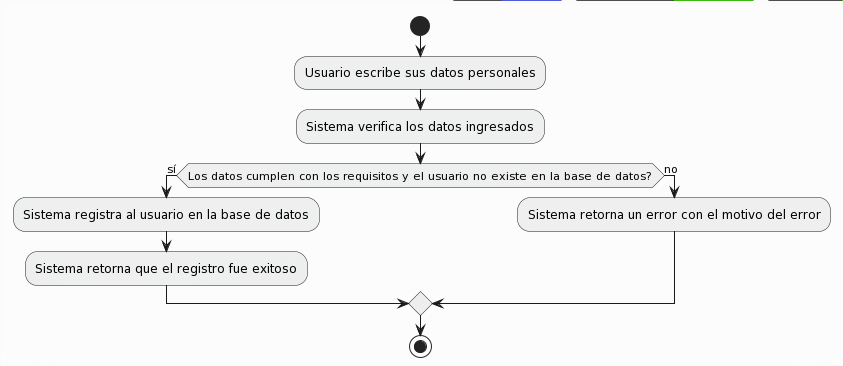
#### Registro de Usuario

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.

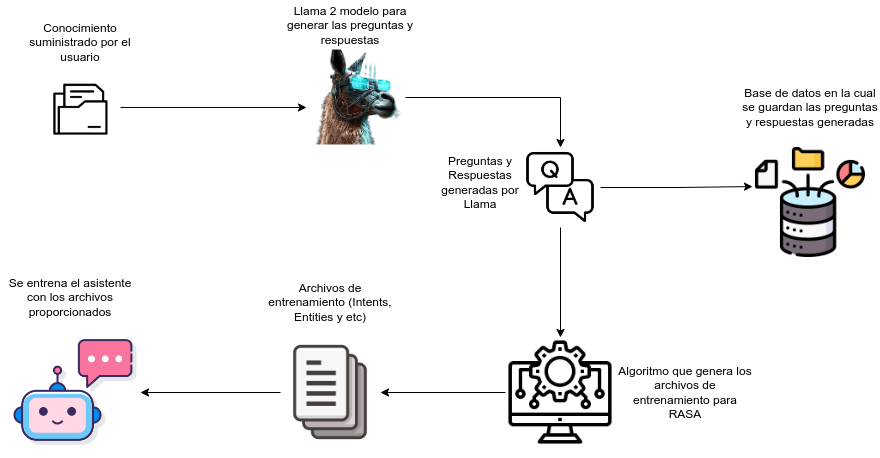
##### Diagrama de Secuencia



##### 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.

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.

## 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 sus métricas lo demuestran.

Si comparamos a Llama 2 con Falcon, como bien se observa en la siguiente foto, en el caso de los modelos de 7Billones 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.

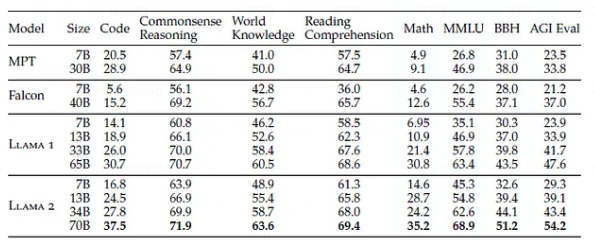


Tabla comparativa entre los modelos de Falcon y Llama [2]

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.

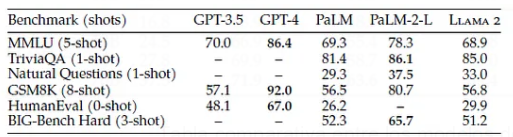


Tabla comparativa entre Llama2, PaLM, PaLM2, GPT-3.5 y GPT-4 [2]

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 en un notebook de Google Colab, 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.

Mencinar las limitantes que tiene Llama como producto OpenSource en cuando a su licencia y el tipo de proyectos en el que se usa.

# Referencias Bibliográficas

[1]shadcn, “Introduction.” Accessed: Oct. 23, 2023. [Online]. Available: https://ui.shadcn.com/docs  
   
   
 [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