# Posibles Flujos en el sistema:

1. Flujo de recolección de datos históricos (Formación de la base de conocimiento).
2. Flujo de captura de preguntas y respuestas en tiempo real en Discord (Con el chatbot en producción).
3. Flujo de una consulta en el chatbot.
4. Flujo de validación de respuestas en Discord.

## Flujo de recolección de datos históricos (Formación de la base de conocimiento)

No se procesa mensaje por mensaje en tiempo real, sino que se extrae todo el historial de un canal para estructurarlo y construir la base de conocimiento desde cero:

1. Se usa la API de Discord para extraer todos los mensajes del canal desde enero 2024 hasta el momento actual.
2. Se pasa cada mensaje a Mistral 7B para determinar: ¿Es pregunta o respuesta? Mistral utiliza un prompt específico para ello y también descarta mensajes irrelevantes como saludos, stickers, memes o mensajes vacíos.
   * + Si es pregunta, se almacena en la base de datos relacional.
     + Si es una respuesta, se almacena en la base de datos relacional y se intenta determinar a qué pregunta pertenece. Para eso, se seleccionan las últimas 5 preguntas almacenadas en la base de datos relacional (ordenadas por fecha anterior a la respuesta) y se le pasa ese conjunto a Mistral junto con el mensaje actual.

Si no se encuentra coincidencia, se prueban las siguientes 5 preguntas anteriores.

Si se determina a qué pregunta pertenece:

* + - * Si es la primera respuesta a esa pregunta, se genera el embedding de la pregunta y se guarda en ChromaDB. Esto se implementa mediante un trigger en la base de datos relacional.
      * Si no es la primera, se recupera el ID del embedding existente de esa pregunta desde ChromaDB para vincularla correctamente.

## Flujo de captura de preguntas y respuestas en tiempo real en Discord (Con el chatbot en producción)

1. Alguien escribe en el canal de “consultas” en Discord
   * Se captura el mensaje usando la API de Discord o discord.py.
2. Se pasa el mensaje a Mistral 7B para clasificarlo:
   * Si es una pregunta: Se almacena en la base de datos relacional con los fragmentos necesarios.
   * Si es una respuesta: Hay que encontrar a qué pregunta pertenece.

* Se toman las últimas 5 preguntas y se pregunta a Mistral si alguna es la correspondiente.
  + Si no hay coincidencia, se prueban otras 5 preguntas.
* Si Mistral encuentra la pregunta correcta:
  + Si se trata de la primer respuesta a una pregunta, se pasa la pregunta a embeddings.
  + En caso contrario (es la segunda, la tercera, etcétera):
    - Se busca la representación de la primer pregunta en la base de datos de vectores (ChromaDB).
    - Se obtiene su id\_chroma.
    - Se busca ese id\_chroma en la tabla de embeddings en la base relacional.
    - Se encuentra su fragmento de pregunta y, con él, la pregunta completa.
    - Se almacena la respuesta vinculada a esa pregunta en la base de datos.
      * La respuesta se divide en fragmentos si supera cierta longitud y se almacena con su estado (validada o no).

### Diferencias clave entre captura en tiempo real y formación de base de conocimiento

| Aspecto | Captura en tiempo real | Historial de 2024 (Base de conocimiento) |
| --- | --- | --- |
| Cómo llegan los mensajes | Se capturan en tiempo real desde Discord | Se extraen en bloques de mensajes desde la API de Discord |
| Objetivo | * Almacenar la pregunta en la base de datos relacional siempre que sea nueva. * Almacenar las respuestas a una pregunta en una base de datos relacional   + Si es la primera respuesta a una pregunta, se guarda dicha respuesta en la base de datos relacional y se generan los embeddings de la pregunta. | * Crear una base de datos relacional para las preguntas, respuestas y sus relaciones. * Crear base de datos de vectores del historial de preguntas. |
| Uso de embeddings | Para la pregunta que recibe su primer respuesta | Para todas las preguntas que tienen al menos una respuesta |
| Uso de Mistral 7B | Se usa para clasificar y relacionar cada mensaje nuevo | Se usa para clasificar y relacionar cada mensaje en el historial |
| Relaciones | Se establece relación pregunta-respuesta en tiempo real | Se reconstruyen relaciones en base a datos pasados. |

|  |  |  |
| --- | --- | --- |
|  |  |  |

## Flujo de una consulta en el chatbot

1. Se captura la pregunta del usuario en Discord.
2. Se genera un embedding de la pregunta y se busca en ChromaDB (comparando con embeddings de preguntas previas).
3. Se obtiene la pregunta más similar y se recuperan sus respuestas desde la base relacional.

* Si hay muchas respuestas:
* Se muestran primero las respuestas validadas con más votos.
* Si no hay respuestas validadas, se muestran respuestas con una advertencia.
* Mistral es quien combina respuestas usando todas las respuestas existentes.
* Si no hay respuestas, se indica que debe realizarse la pregunta en Discord porque nunca antes se hizo.

1. Se muestra en el chatbot la respuesta correspondiente.

## Flujo de validaciones de respuestas en Discord

1. Hay captura de reacciones en Discord. Se usa la API de Discord para detectar reacciones.
   * Se analiza si se trata de reacción para una respuesta, caso contrario no se continúa el flujo.
2. Se analiza la reacción específica: si es de un docente y cuál es.
   * Si una respuesta recibe una reacción de validación de un docente, se marca como validada en la base de datos relacional (campo validada = True en la tabla respuestas y se actualiza cantidad\_validaciones)
   * Si no tiene reacciones del docente, sigue sin validación.

Nota: Cuando un alumno realiza una consulta en el chatbot, en la búsqueda de respuestas, si hay muchas, se ordenan por número de validaciones para combinarse y mostrarse.