

Inteligencia artificial y programación

Servicio tipo “ChatGPT” local con Serge-Chat y Mistral (Gratis y privado)



Autor: Sergi García

Actualizado Enero 2024




Licencia



Reconocimiento - No comercial - CompartirIgual (BY-NC-SA): No se permite un uso comercial de la obra original ni de las posibles obras derivadas, la distribución de las cuales se ha de hacer con una licencia igual a la que regula la obra original.

Nomenclatura

A lo largo de este tema se utilizarán diferentes símbolos para distinguir elementos importantes dentro del contenido. Estos símbolos son:

 **Importante**

 **Atención**

 **Interesante**

ÍNDICE

1. Introducción	3
1.1 ¿Qué software utilizaremos?	3
2. Instalando Serge-Chat a través de Docker	3
3. Obteniendo modelo "Mistral" para "Serge-Chat"	4
4. Añadiendo modelos no disponibles en el menú a Serge-Chat	6
5. Otras alternativas a "Serge-chat": GPT4All	7

SERVICIO TIPO "CHATGPT" LOCAL CON SERGE-CHAT Y MISTRAL (GRATIS Y PRIVADO)

1. INTRODUCCIÓN

En el panorama actual de las interacciones digitales, la demanda de servicios de chat con inteligencia artificial, similares a "ChatGPT". "ChatGPT" es una interesante y potente herramienta.

Sin embargo, surge un dilema tanto con "ChatGPT" como con otras soluciones existentes: el costo de adquisición/suscripción y la preocupación por la privacidad (envías información, que puede ser sensible a un tercero, como código, datos personales, estrategias empresariales, etc.).

Para abordar estas inquietudes, en este documento ofrecemos una alternativa gratuita y "self-hosted" para garantizar la máxima privacidad, utilizando Serge-Chat y el modelo "Mistral".

! Atención: los requisitos mínimos pueden variar según el modelo utilizado. Por lo general siguen esta norma, según los billones (americanos) de parámetros: 7b -> 8GB de RAM, 13b -> 16 GB de RAM y 43b -> 32 GB de RAM

1.1 ¿Qué software utilizaremos?

Para esta tarea, en resumen, utilizaremos el siguiente software:

- **Serge-Chat:** <https://github.com/serge-chat/serge>
 - Software para descargar, preparar y lanzar distintos modelos LLMs e interactuar con ellos con una interfaz similar a "ChatGPT" vía web.
 - También para usos más avanzados, posee una API Rest.
- **Mistral:** el modelo "Mistral" para Llama. Su web oficial <https://mistral.ai/>

2. INSTALANDO SERGE-CHAT A TRAVÉS DE DOCKER

Si no tienes instalado Docker o no sabes utilizarlo, puedes encontrar más información en mi curso

<https://github.com/sergarb1/CursoIntroduccionADocker>

Aunque el código de Serge-Chat está disponible y se puede poner en marcha sin dockerizar, la recomendación del propio autor es lanzarlo de forma dockerizada.

El propio autor en su web, indican tanto un comando Docker para ponerlo en marcha, como una propuesta de fichero "docker-compose.yml".

Esta información disponible en: <https://github.com/serge-chat/serge>

Siguiendo esta guía, para instalar Serge-Chat dockerizado, nos basta con este comando (todo en una sola línea) el siguiente comando:

```
docker run -d --name serge -v weights:/usr/src/app/weights -v  
datadb:/data/db/ -p 8008:8008 ghcr.io/serge-chat/serge:latest
```

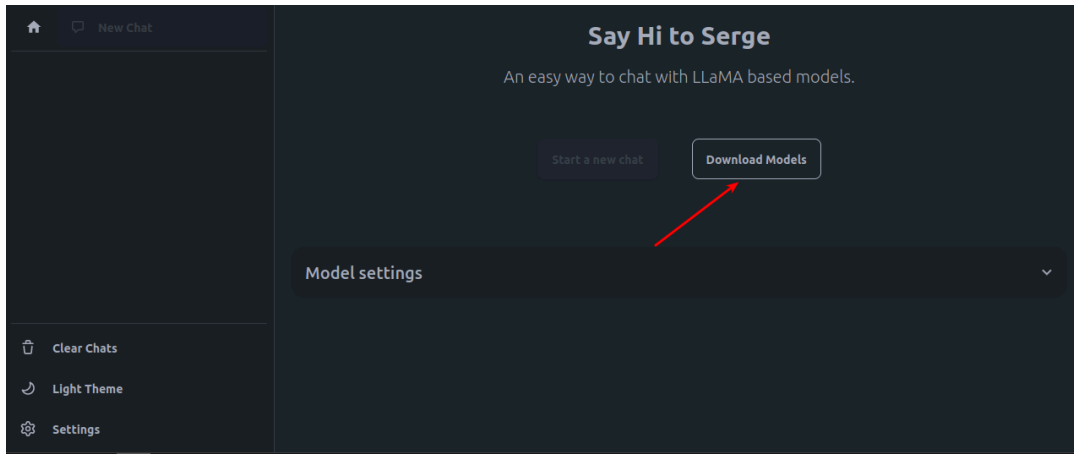
Con ello, tendremos funcionando Serge-Chat en el puerto 8008. Podemos acceder para probar su correcto funcionamiento en local mediante <http://localhost:8008> y si queremos tener información sobre su API, la podemos encontrar en <http://localhost:8008/api/docs>.

Asimismo, si queremos acceder remotamente, simplemente cambiaremos "localhost" en nuestro navegador por el host o dirección IP adecuada.

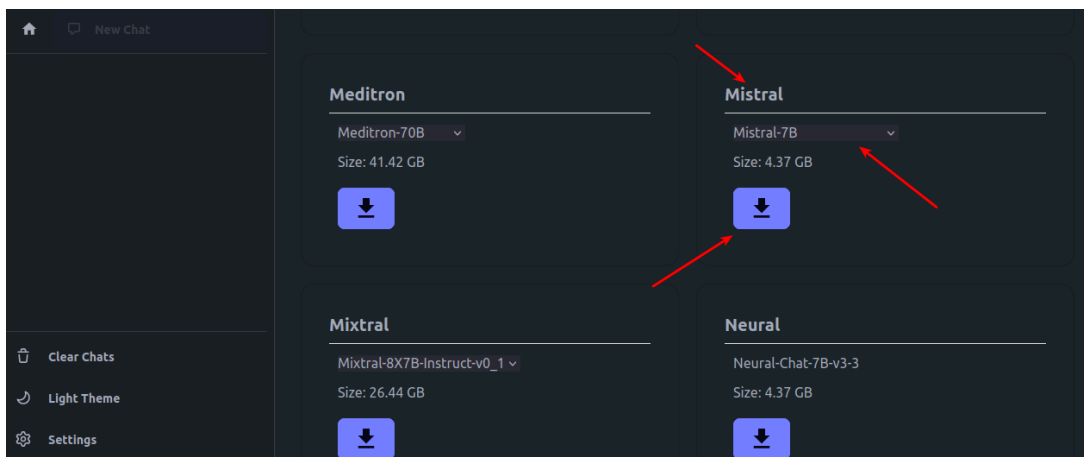
3. OBTENIENDO MODELO "MISTRAL" PARA "SERGE-CHAT"

En este caso, simplemente, acudiremos a "Serge-Chat" mediante <http://localhost:8008> (o el host/IP que proceda) y descargaremos el modelo mediante la interfaz web, de una forma similar a como se ve en las siguientes imágenes:

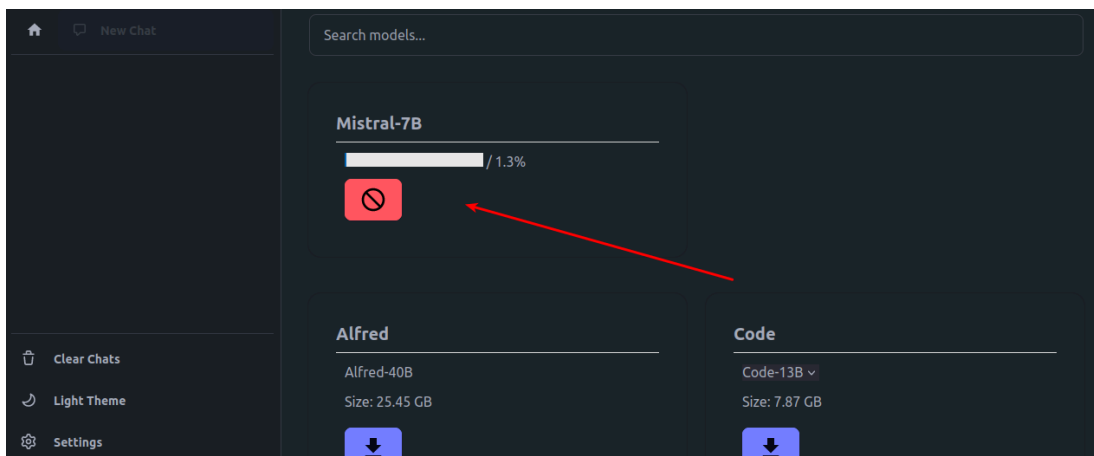
1) Vamos a la lista de modelos disponibles. Podemos ir siguiendo lo indicado en la imagen, o acudir directamente vía URL <http://localhost:8008/models>.



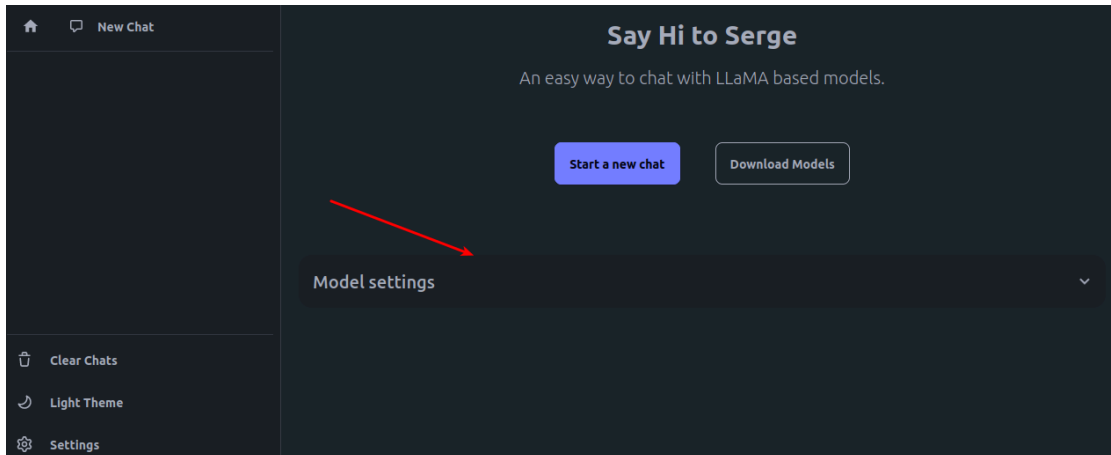
2) Tras ello, elegimos el modelo "Mistral" y dentro del desplegable, elegimos "Mistral 7B".



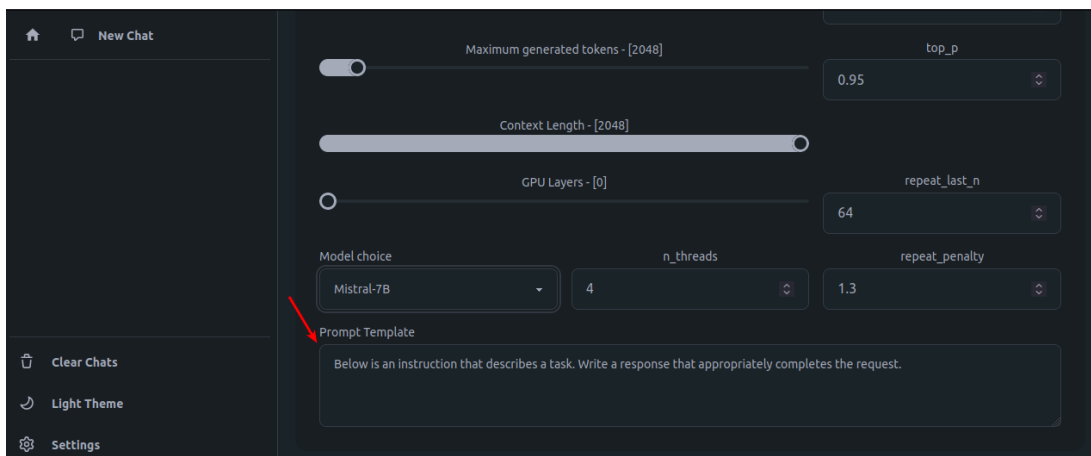
3) Tras ello, si subimos el "scroll" hacia arriba, veremos que ha comenzado el proceso de descarga.



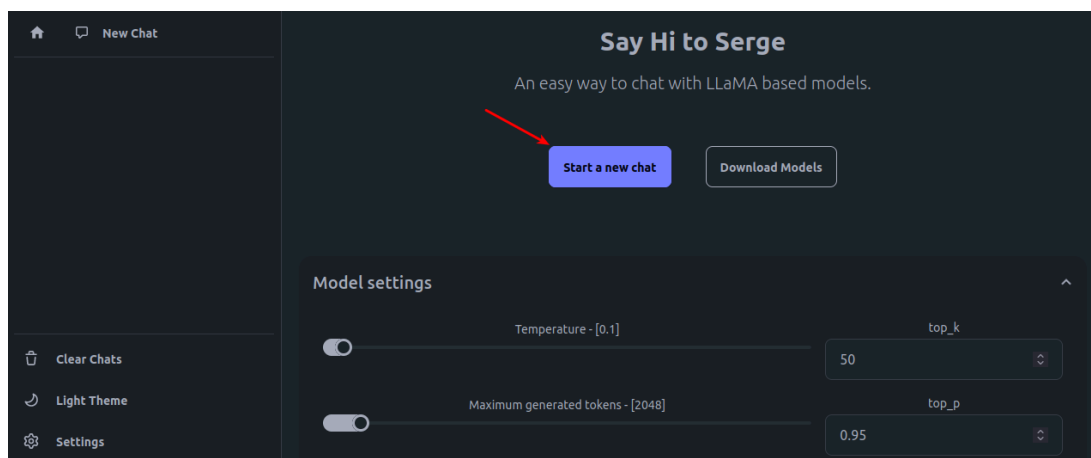
4) Cuando haya finalizado, ya podemos utilizar Serge-Chat para conversar con el modelo con un estilo similar a "ChatGPT". Pero antes de comenzar nuestro chat, configuraremos (de forma opcional) algunas opciones del modelo. Accedemos a ella de esta forma:



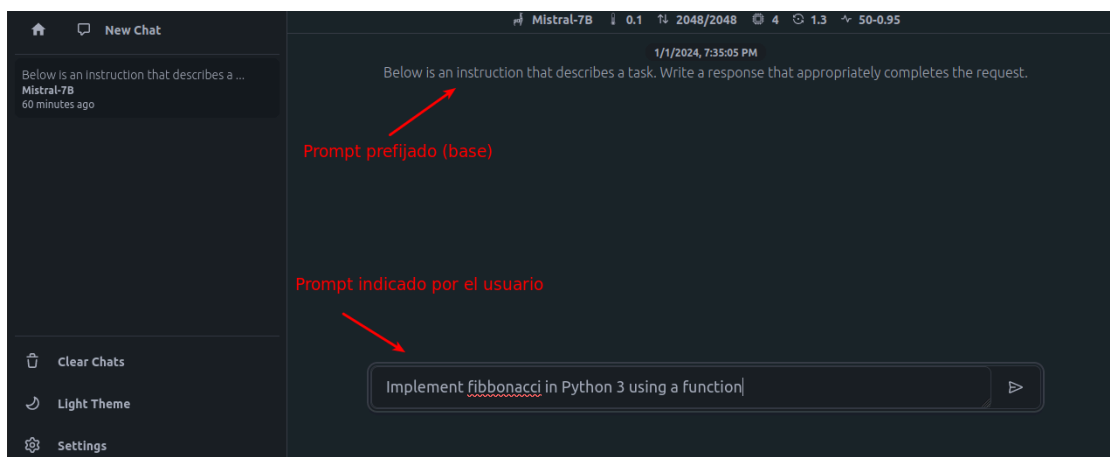
5) Una vez abierto el desplegable "Model settings" podemos configurar algunos parámetros del modelo, incluyendo un "prompt base" que utilizará como rol para nuestro chat y podemos modificar a nuestra conveniencia.



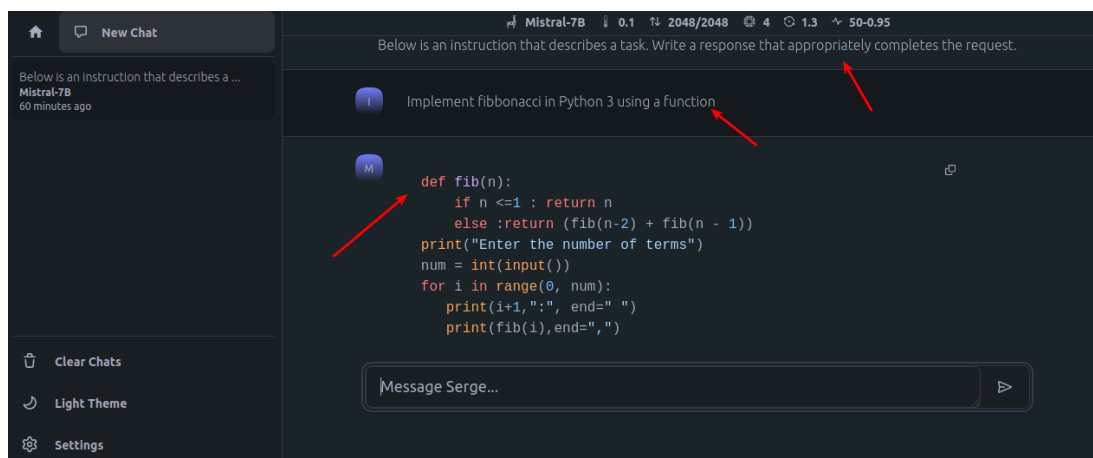
6) Una vez todo listo, volveremos arriba y pulsaremos en "Start a new chat"



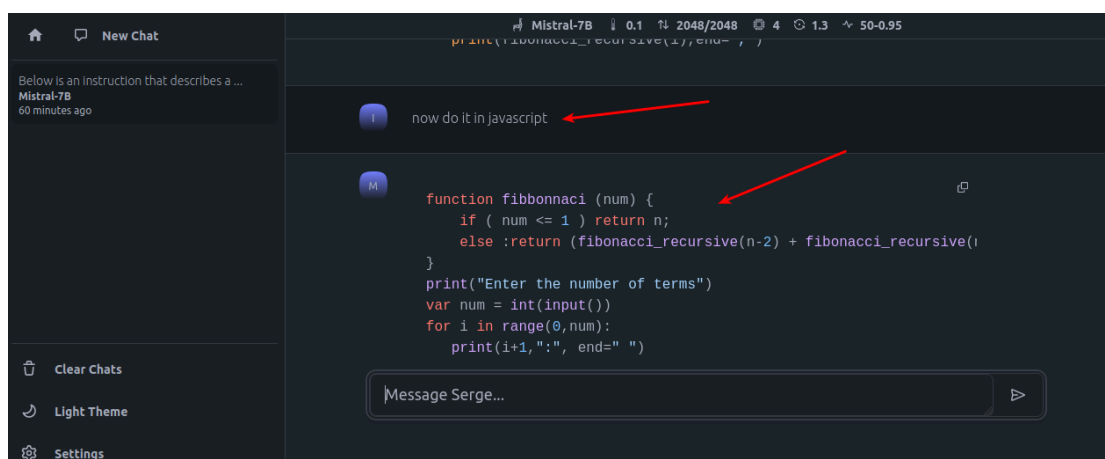
7) Una vez en el chat, escribiremos nuestro prompt y esperaremos respuesta. Importante recordar que lo indicado arriba, es un prompt "prefijado" que podemos cambiar en la configuración del modelo. La parte inferior es mi texto para la conversación.



8) Aquí vemos el resultado de la interacción:



9) Si lo deseamos, podemos seguir manteniendo la conversación, como se ve en este ejemplo:



4. AÑADIENDO MODELOS NO DISPONIBLES EN EL MENÚ A SERGE-CHAT

En el caso de querer utilizar un modelo no disponible para descarga dentro de los modelos de Serge-Chat, no hay problema, ya que se puede añadir cualquier modelo "custom". Para ello, debes obtener el modelo en formato GGUF, copie el archivo en la carpeta "weights" dentro del contenedor Docker de Serge-Chat y finalmente, cambiar la extensión del modelo copiado a ".bin". Esto puede ser útil, por ejemplo, para incluir modelos más adaptados a determinados idiomas,

como LANCE Mistral, <https://huggingface.co/clipbrain/lance-mistral-7b-it-es> que refina el modelo "Mistral 7b" para que funcione mejor con instrucciones en castellano.

5. OTRAS ALTERNATIVAS A "SERGE-CHAT": GPT4ALL

Aunque inicialmente más popular, ya que llego antes, a título personal me gusta más "Serge-Chat", lo que no quita que GPT4All sea una gran alternativa a este software, ya que tienen funciones similares. Una de las principales pegadas de GPT4All es que a día de hoy está orientado a la instalación y no a la dockerización (solo hay soluciones de terceros y amenudo no actualizadas).

Si quieres instalar y descargar GPT4All, puedes hacerlo en <https://gpt4all.io/> y ahí seguir las instrucciones de instalación según tu sistema operativo.

Por ejemplo, en sistemas Linux descarga el instalador llamado "gpt4all-installer-linux.run" y ejecutalo como "root". Tras ello, sigue los pasos de la instalación.

```
./gpt4all-installer-linux.run
```

La forma de trabajar es similar a "Serge-Chat", así que una vez instalado te recomendamos explorar y antes dudas visitar su web <https://gpt4all.io/> y su repositorio de GitHub <https://github.com/nomic-ai/gpt4all>.