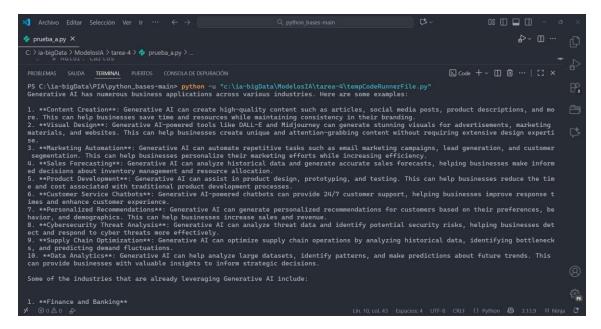
## PRUEBA\_A

En prueba a.py hemos conectado Python con el servidor local de Ollama



¿Qué diferencia hay entre 'stream' true y false? 

Stream true, el modelo va escribiendo poco a poco, letra por letra, sin embargo en false el modelo primero manda todo el contenido y luego se imprime

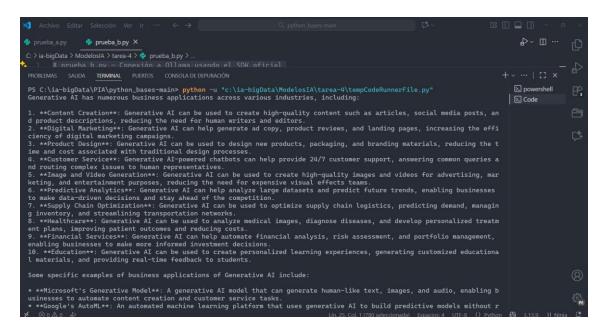
¿Dónde vive el texto de salida en la respuesta JSON? 

Dentro de la variable response[mensaje][content]

¿Qué HTTP status esperarías en éxito? → 200

## PRUEBA\_B

Nos hemos comunicado con el modelo de Ollama usando su SDK



¿Qué ventajas e inconvenientes ves entre usar la API HTTP y el SDK? DEN cuanto a código es bastante más corto utilizando el SKD sin embargo en cuanto a velocidad es más o menos la misma

¿Qué campos serían útiles en 'options' para controlar el estilo de salida? -> Alguna opción para formatear el mensaje o generártelo en un pdf.

¿Cómo validarías que el modelo realmente corresponde a 'llama3.2' y no a otro? > Fijándote en el modelo que pone en la variable que es el que va a utilizar Ollama al ejecutar el script

Si quisieras internacionalizar el prompt, ¿qué cambios harías? → Crear de alguna forma algún método capaz de adaptar la respuesta al idioma del usuario.