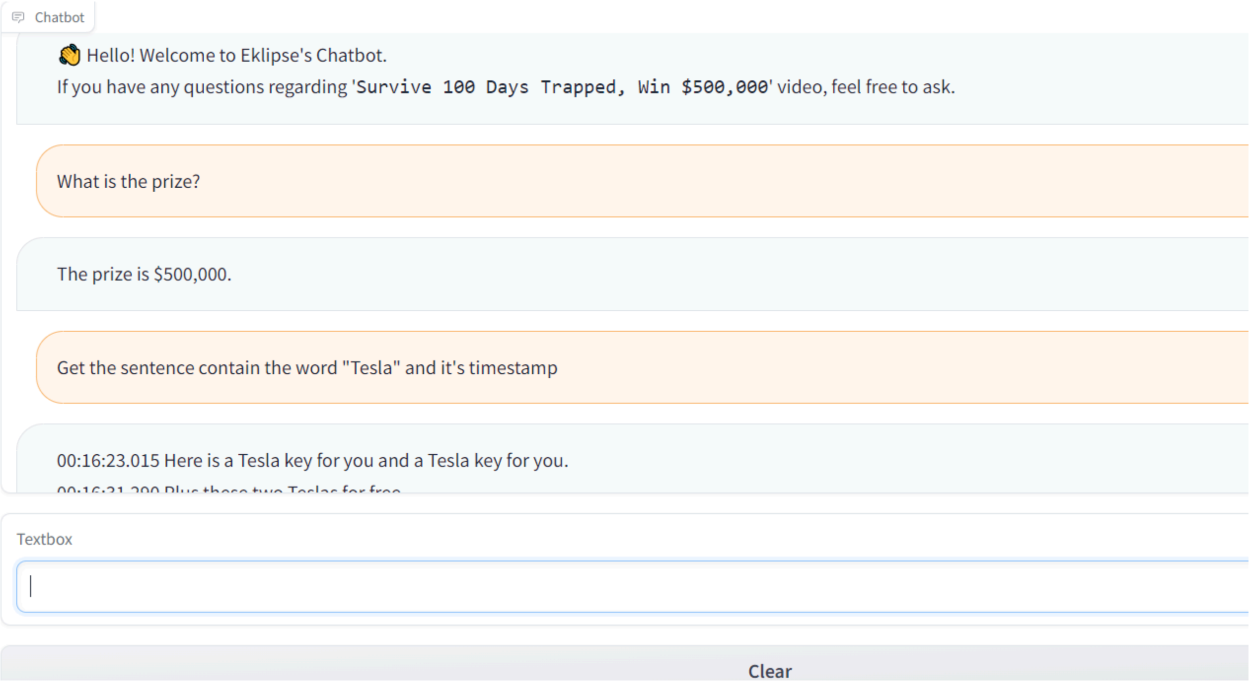
Exercise 1 - Chatbot for Youtube video

To enhance user experience by providing a tool for them to search any specific moments in a video, we will develop a chatbot. The chatbot will serve as an assistant allowing the users to ask anything about content insight in the videos. For example, the following image from the eklipse's chatbot that allows the users asking information from the content of the following video https://www.youtube.com/watch?v=9RhWXPcKBI8:



👋 Hello! Welcome to Eklipse's Chatbot.
If you have any questions regarding 'Survive 100 Days Trapped, Win $500,000' video, feel free to ask.

What is the prize?

The prize is $500,000.

Get the sentence contain the word "Tesla" and it's timestamp

00:16:23.015 Here is a Tesla key for you and a Tesla key for you.
00:16:31.200 Plus these two Teslas for free

In this exercise, your task is to propose a pipeline in detail and build a chatbot with the input are mp4 videos (or Youtube links) and the output that can answer any question about the content of those videos properly. You may utilize any closed or open-source models or frameworks.

Requirements:

- Provide a brief report of your pipeline and develop an application with input as a custom mp4 video (or Youtube link).
- The chatbot can support image(s) as the prompting input would be highly appreciated. For example, a user can send a picture of a person and then ask the chatbox about the timestamp of the person appearing in the video.

Exercise 2 - LLM deployment (Optional)

To reduce costs of using ChatGPT, we aim to develop an on-premises LLM. Your task is to find a solution for deploying an open-source LLM efficiently. There are several factors should be considered to make the on-premise LLM more efficient:

- VRAM: The amount of GPU memory in GB required when deploying a model.
- Speed: The metric should be average completion tokens per second.

- Performance: The model's accuracy and consistency compared to the original model.
- Feature Support (optional): In addition to token completions, we also want to have the log probabilities of those tokens.

In this exercise, you should propose a solution to deploy model BLOOMZ 1b1 https://huggingface.co/bigscience/bloomz-1b1. This task is designed to test your LLM knowledge and deployment skills, …

Requirements:

- Provide a brief report of your solution and your API Restful Docker (or code). Any input and output formats are allowed. We will evaluate your Docker (or code) on our server.
- Report exactly metrics (VRAM, Speed and Performance) and technique you use in this exercise.
- A README.md file is very useful in this case. It would be nice to include a README to guide on how to run your Docker (or code).