8th December 2024

AI Conversational ChatBot Project

About:

I created an AI Conversational ChatBot that operates with Cerebras' Inference API designed to provide quick, smart and live replies across various virtual spaces. The chatbot takes input from the user and supplies the proper response, explaining how Cerebras can improve output and optimize scope in AI-related applications.

How I built it:

The chatbot was created with Node.js for the backend with webhooks to handle messages and with Ngrok to share the local server for API integration. The Cerebras Inference API was the main tool used during the project because it allowed quick and efficient generation of messages and delivery of responses through AI.

How the project leverages Cerebras' fast inference:

The Inference API developed by Cerebras proved to be very useful in this project due to its low latency and fast result response time. On natural language models, it was able to perform complex computations with ease thus enhancing the capabilities of the chatbot to provide accurate responses without latency. Thanks to Cerebras' scalability and performance, I came up with a solution that was able to sustain a high level of interactivity while having to handle a lot of requests at once.

Future RoadMap:

- Performance Optimization: Scale up the system as a real-world deployment for these high-demand applications by utilizing Cerebras' infrastructure as a stable and fast system.
- Project Enhancement:
  - Expand on the conversational context and consider extending the user interactions and the Internet trends identified in this paper, using Cerebras' AI applications for analyzing the users' preferences.
  - Improve the context of the conversation along with analyzing user preferences using Cerebras APIs.