ChatGPT 2.0 Local

Electiva II

Docente a cargo

Cristhian Alejandro Cañar Muñoz

Presentado por

Ulises ortega revelo

Universidad autónoma del cauca-2025

Resumen: Se desarrolló una aplicación web de chat basada en modelos de lenguaje grandes (LLMs), completamente funcional en el navegador y sin conexión a internet. Se implementó utilizando HTML, CSS y JavaScript, apoyándose en la librería WebLLM y la API WebGPU para ejecutar localmente modelos de lenguaje como Llama-3-8B-Instruct y RedPajama-INCITE-Chat-3B, priorizando privacidad y eficiencia.

Abstract: A fully functional, browser-based, offline-enabled chat web application based on large language models (LLMs) was developed. It was implemented using HTML, CSS, and JavaScript, leveraging the WebLLM library and the WebGPU API to locally execute language models such as Llama-3-8B-Instruct and RedPajama-INCITE-Chat-3B, prioritizing privacy and efficiency.

I. INTRODUCTION

Con el avance de la inteligencia artificial, los modelos de lenguaje han cobrado gran relevancia por su capacidad para generar texto, asistir al usuario y automatizar tareas. Generalmente, estos modelos requieren infraestructura potente y conexión constante. Este proyecto explora una solución completamente local, que mediante tecnologías web modernas permite ejecutar modelos LLM en el navegador, eliminando la dependencia del servidor y aumentando la privacidad del usuario.

II. MARCO TEORICO

Modelos de Lenguaje Grandes (LLMs): Son redes neuronales entrenadas con grandes volúmenes de texto que permiten entender, generar y traducir lenguaje natural. Entre los modelos utilizados en este proyecto están:

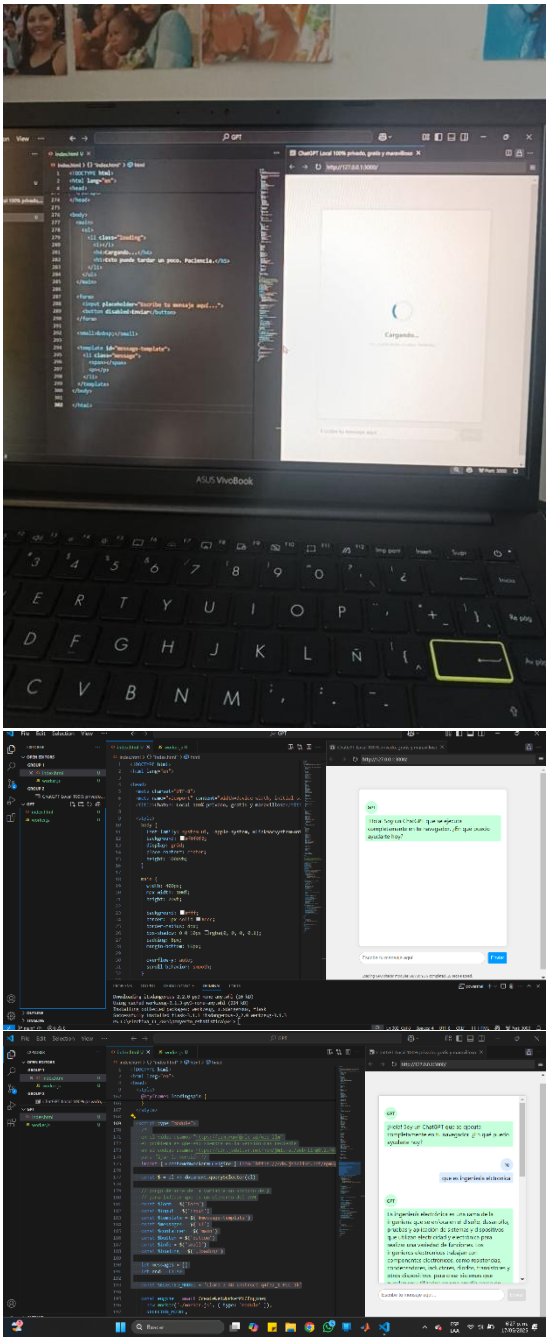
- LLaMA-3-8B-Instruct: Modelo de 8 mil millones de parámetros, diseñado para tareas de instrucción con alto desempeño, aunque con una carga computacional elevada.
- RedPajama-INCITE-Chat-3B: Modelo de 3 mil millones de parámetros optimizado para tareas conversacionales, con menor requerimiento de recursos y buen rendimiento en navegadores con soporte WebGPU.

WebLLM:
Es una biblioteca que permite ejecutar modelos LLM directamente en el navegador. Está construida sobre MLC (Machine Learning Compilation) y aprovecha tecnologías como WebGPU para compilar y ejecutar modelos eficientemente sin necesidad de servidor.

WebGPU:
Es la nueva API de gráficos del navegador que permite acceder directamente a la GPU del dispositivo. Sustituye a WebGL y mejora notablemente el rendimiento en tareas de computación intensiva, como el procesamiento de modelos de IA.

Tecnologías Web (HTML, CSS, JavaScript): La aplicación fue desarrollada con tecnologías estándar de la web. Se utilizó JavaScript para la lógica de interacción, HTML para la estructura y CSS para el estilo visual del entorno de chat.

III. PRUEBAS



IV. CONCLUSIONES

La implementación de modelos de lenguaje grandes directamente en el navegador representa un avance significativo en cuanto a privacidad, accesibilidad y portabilidad de aplicaciones basadas en inteligencia artificial. A través de este proyecto, se comprobó que es posible ejecutar modelos como LLaMA-3-8B-Instruct y RedPajama-INCITE-Chat-3B sin conexión a internet, mediante la tecnología WebLLM y el soporte de WebGPU.

Sin embargo, también se identificaron importantes limitaciones técnicas. Durante las pruebas, se observó un alto consumo de recursos computacionales, tanto de CPU como GPU, lo que provocó lentitud en la carga inicial y en algunos momentos del procesamiento. Especialmente con modelos más pesados como LLaMA-3, la experiencia fue notablemente más lenta. Al cambiar al modelo RedPajama, más ligero, se obtuvo un mejor equilibrio entre velocidad y calidad de las respuestas.

A pesar de la demanda técnica, las respuestas generadas fueron coherentes, contextuales y útiles, demostrando la efectividad de los modelos aún en entornos limitados como el navegador.

Este experimento demuestra que, aunque aún no es ideal para dispositivos de bajo rendimiento, el futuro de la IA local y privada en el navegador es prometedor. Con la evolución de WebGPU y la optimización de modelos, este enfoque podrá escalar hacia aplicaciones más ligeras, accesibles y seguras para todos los usuarios.