Desarrollo de un Chatbot de Asistencia al Cliente con Fine-Tuning de LLM

---

1. Problema a Resolver

- **Objetivo:** Crear un chatbot en español capaz de responder preguntas frecuentes de clientes (por ejemplo, horarios de atención, políticas de devolución, métodos de pago, etc.).
- **Desafío:** Lograr que el modelo genere respuestas coherentes y precisas mediante fine-tuning, utilizando un modelo preentrenado adaptado a un entorno de asistencia al cliente.

---

2. Enfoque y Solución Propuesta

- **Modelo Base:** Se eligió un modelo preentrenado en español (basado en GPT-2) y se cuantizó en 4 bits (QLoRA) para optimizar recursos.
- **Fine-Tuning con Adaptadores (LoRA):** Se adjuntaron adaptadores para hacer que el modelo cuantizado sea entrenable, sin necesidad de actualizar todos los parámetros.
- **Implementación:** Se desarrolló un pipeline de entrenamiento e inferencia en Visual Studio Code, utilizando técnicas de fine-tuning supervisado con un dataset que une la "instrucción" y la "respuesta" en un mismo prompt.

---

3. Complicaciones Encontradas

- **Dataset Muy Pequeño y Poco Diverso:**
    - Solo se utilizaron 11 ejemplos, lo cual es insuficiente para que el modelo aprenda patrones generalizables y cubra la variabilidad en las consultas de clientes.
    - El propósito era utilizar un dataset en español de atención al cliente, pero no logré encontrarlo.
- **Respuestas Incoherentes:**
    - Debido a la limitación en la cantidad y diversidad de datos, el modelo genera respuestas largas y a veces sin sentido, ya que no se han establecido restricciones adecuadas en el proceso de generación.
- **Falta de Evaluación y Refinamiento:**
    - La ausencia de un proceso sistemático de evaluación y ajustes iterativos impide detectar y corregir salidas no deseadas, lo que afecta la coherencia de las respuestas.

---

4. Propuestas de Mejora

- **Ampliar y Diversificar el Dataset:**
    - Incorporar más ejemplos y casos de uso para cubrir una mayor variedad de consultas y respuestas realistas de atención al cliente.
- **Ajuste de Parámetros en la Generación:**
    - Establecer límites en la longitud de las respuestas y afinar los parámetros de sampling (como temperatura, top_k y top_p) para generar respuestas más precisas y coherentes.

- **Implementar Evaluación Iterativa:**
    - Definir métricas de evaluación y realizar pruebas de validación para identificar salidas incoherentes y ajustar el modelo de forma progresiva.

---

5. Conclusión

Aunque se ha implementado un pipeline para crear un chatbot de asistencia al cliente con fine-tuning de un LLM cuantizado, las limitaciones del dataset y la generación sin restricciones han generado respuestas incoherentes. La clave para mejorar es ampliar y diversificar los datos de entrenamiento, junto con la implementación de un proceso de evaluación y ajuste continuo.
