Entrenamiento del modelo LLaMA 3.2 (1B) en el clúster HPC-UCR con QLoRA para adaptar un modelo instructivo en español

RESUMEN

En América Latina, la adaptación de modelos de lenguaje se ve limitada por los altos costos computacionales y la predominancia del inglés.

Este trabajo presenta una estrategia eficiente para adaptar el modelo LLaMA 3.2 (1B) al español mediante QLoRA, demostrando la viabilidad de entrenar modelos instructivos en entornos académicos de alto desempeño (HPC-UCR) de forma sostenible y reproducible.

METODOLOGÍA

- Dataset de ~400 MB (~50 M tokens) de instrucciones en español.
- Entrenamiento total de ~18 h (3 bloques de 6 h) en GPU mediante SLURM con reanudación automática por checkpoints.
- Monitoreo de métricas de loss y perplexity.

RESULTADOS

- Tiempo medio de inferencia: 0.8–1.2 s por instrucción en evaluación.
- Flujo reproducible y adaptable para otros modelos.
- Promueve la soberanía tecnológica y lingüística latinoamericana.
- Reducción progresiva del loss y la perplejidad a lo largo del entrenamiento.

ENTORNO COMPUTACIONAL

- Procesamiento en CPU: 16 nodos Lenovo ThinkSystem SD630 V2, cada uno con 64 cores Intel Xeon Gold 6338 (1028 en total) y 16 GB RAM por core.
- Procesamiento en GPU: 2 nodos Lenovo ThinkSystem SR670 V2, cada uno con 4 GPUs NVIDIA Tensor Core A100 (80 GB), 64 cores Xeon Gold 6338, y 16 GB RAM por core.
- Lenguaje : Desarrollo en Python dentro del entorno conda del HPC-UCR.

LinkedIn

Autora: Alison Lobo Salas
Institución: Universidad de Costa Rica – CIOdD

TICAL 2025 · San José, Costa Rica · 11–13 de noviembre