Ampliación de Fundamentos de Hardware

Angel Fabrizio Ullaguari Yanza

2ºASIR

Práctica IA. Cuantización

Práctica **3**

Práctica

Preparación del entorno. Cargamos el modelo original Phi-3-mini-4k-instruct de Microsoft desde Hugging Face.

Ahora probamos el modelo cuantizado.

Aquí se comparan los dos archivos en la lista de modelos. El modelo original Phi 3 Mini F16 (no cuantizado) ocupa más espacio almacenado que el modelo Q4 (cuantizado).

Verificación y Prueba. Esta captura muestra una comparativa de velocidad de generación de texto con ambos modelos. En nuestro caso nos salió al revés, o sea, el modelo no cuantizado nos da la respuesta con menos tokens por segundo cuando debería ser el modelo cuantizado el que dé menos tok/sec.