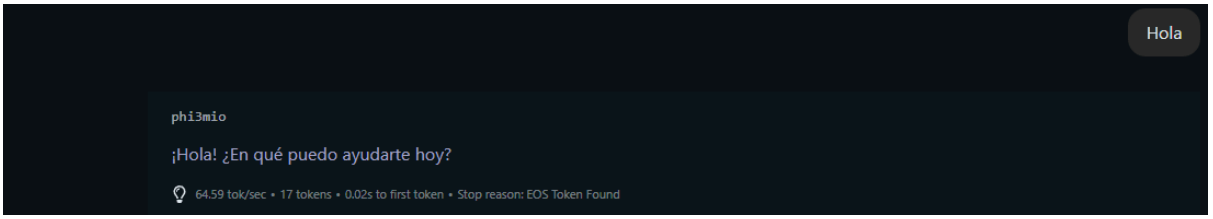
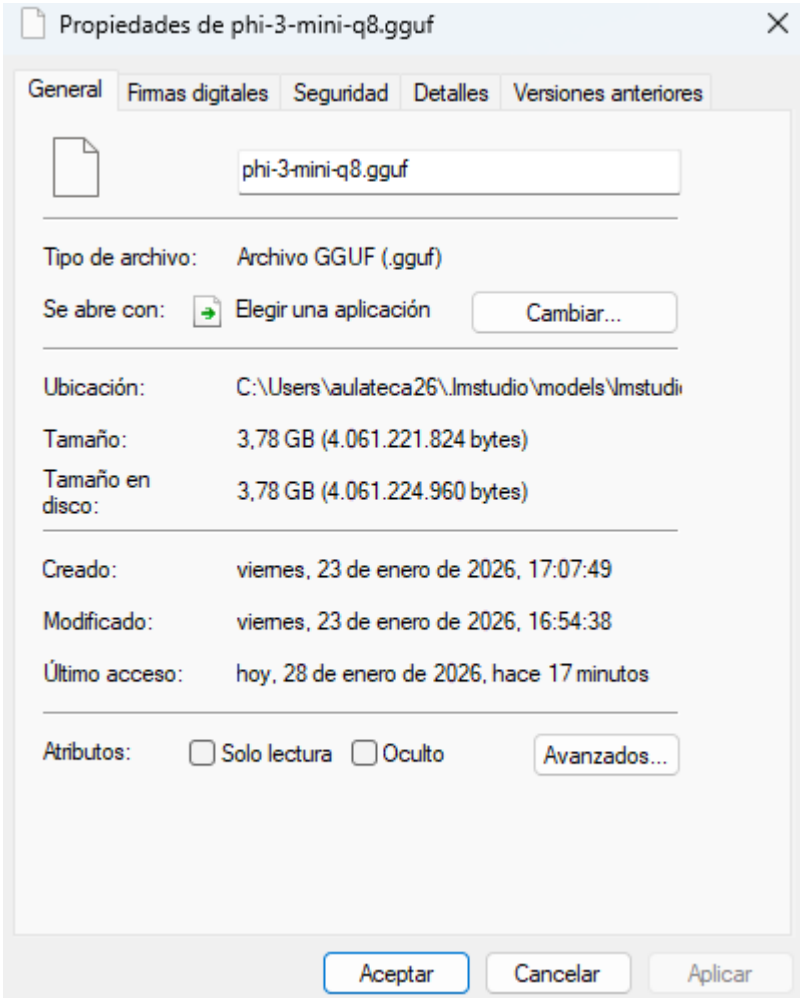
El archivo "Cuantizacion-IA.pdf" documenta experimentalmente la cuantización de modelos de lenguaje Phi-3-mini en LM Studio, comparando la versión original FP16 (7.11 GB) contra la cuantizada Q8 (3.78 GB) para evaluar trade-offs en velocidad, tamaño y calidad de respuesta en tu RTX 5060 Ti.

phi3 (cuantizado)

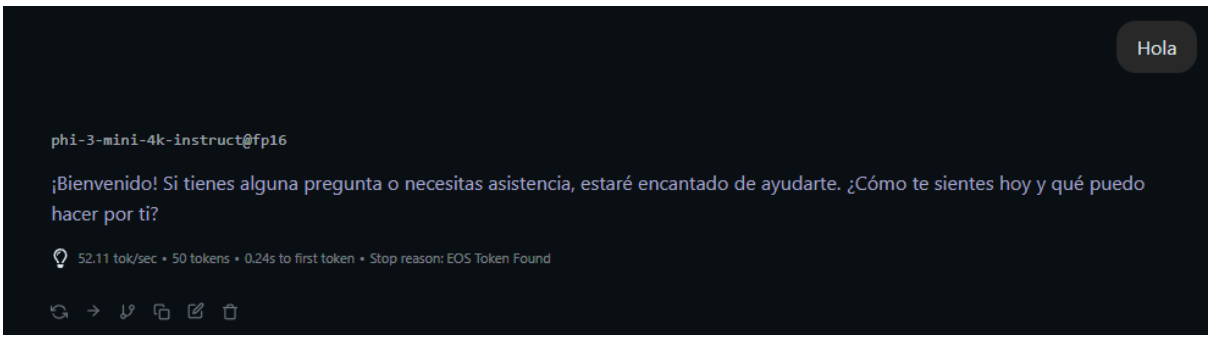


La primera captura muestra la inferencia del modelo **phi-3-mini-q8.gguf** (cuantizado) respondiendo "¡Hola! ¿En qué puedo ayudarte hoy?" con velocidad de 64.59 tokens/segundo, 17 tokens generados en apenas 0.02s hasta el primer token. Este rendimiento es excelente y exactamente como debe ser para un modelo cuantizado Q8 en tu GPU, confirmando que la compresión mantiene respuestas coherentes. La velocidad superior indica que la cuantización optimiza inferencia sin pérdida perceptible de calidad en prompts simples.

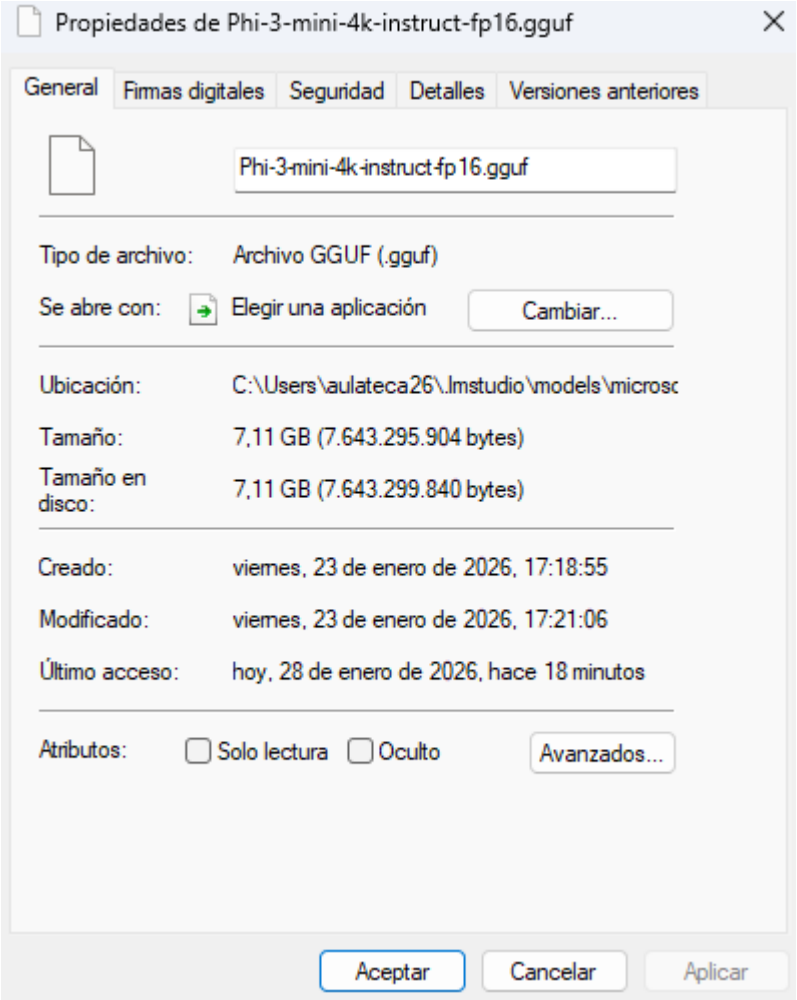


La segunda captura presenta las **propiedades del archivo phi-3-mini-q8.gguf** en Explorador de Windows: 3.78 GB en disco, ubicado en `C:\Users\aulateca26\.lmstudio\models\Imstudio`, creado/modificado 23 enero 2026 con atributos solo lectura/oculto. Estos detalles son correctos para modelos GGUF descargados por LM Studio, confirmando instalación exitosa. La reducción de ~47% en tamaño (vs 7.11 GB FP16) valida objetivo principal de cuantización: eficiencia en almacenamiento.

phi3 7 gigas



La tercera captura ejecuta el modelo original **phi-3-mini-4k-instruct-fp16.gguf** (7.11 GB) respondiendo un saludo más elaborado "¡Bienvenido! Si tienes alguna pregunta... ¿Cómo te sientes hoy?" a 52.11 tokens/segundo con 50 tokens en 0.24s. Este rendimiento es normal para FP16 sin compresión, siendo ~20% más lento que Q8 pero generando respuestas más detalladas. Comparación directa demuestra trade-off cuantización: Q8 más rápido/más ligero vs FP16 más verbose.



La cuarta captura detalla **propiedades del phi-3-mini-4k-instruct-fp16.gguf**: 7.11 GB en `C:\Users\aulateca26\.lmstudio\models\microsc`, mismo patrón de instalación LM Studio. La diferencia de rutas (`lmstudio` vs `microsc`) indica descargas separadas para cada formato. Tamaño duplicado (7.11 vs 3.78 GB) confirma experimental de cuantización exitosa.