La elección entre TinyLlama y Phi-3 para el manejo de modelos de lenguaje grande (LLM) con FastAPI depende en gran medida de tus necesidades específicas y las limitaciones de tu entorno.

TinyLlama es una biblioteca que permite cargar y utilizar modelos de lenguaje grande de manera eficiente. Es útil si tienes restricciones de memoria, ya que permite cargar solo una parte del modelo en la memoria a la vez.

Por otro lado, Phi-3 es un marco de trabajo para la construcción de aplicaciones de procesamiento de lenguaje natural (NLP) que incluye soporte para modelos de lenguaje grande. Es una buena opción si estás construyendo una aplicación más compleja que necesita más que solo la funcionalidad básica de carga y generación de modelos.

Ambas opciones son compatibles con FastAPI, por lo que puedes elegir la que mejor se adapte a tus necesidades. Si solo necesitas cargar y generar a partir de un modelo de lenguaje grande y la memoria no es un problema, Phi-3 podría ser una opción más sencilla. Si la memoria es una preocupación, entonces TinyLlama podría ser la mejor opción.