1. **Phi-3.5-mini-instruct**

For the fine-tuning project, we selected "Phi-3.5-mini-instruct" as the primary model and "gemma-2-27b-bnb-4bit" as the secondary model. The choice of "Phi-3.5-mini-instruct" was made due to its balance between computational efficiency and performance. This model is optimized for instruction-following tasks and provides a good trade-off between speed and resource usage, which enables quick experimentation and iteration without requiring extensive computational resources. This makes it a practical choice for our fine-tuning workflow while still ensuring relevant outcomes in training sessions.

1. **gemma-2-27b-bnb-4bit**

We selected "gemma-2-27b-bnb-4bit" as the secondary model because of its large scale and advanced language capabilities. With 27 billion parameters and 4-bit quantization support, this model offers superior language understanding and generative capabilities. It is well-suited for tasks requiring high-level reasoning, complex interactions, and deep language comprehension. By integrating this secondary model, we can achieve better performance and richer contextual understanding, which is crucial for applications requiring higher fidelity outputs and robust interactions.

**Why decision**

The goal of this project is to fine-tune the model as an English-to-Arabic translator, which is an essential tool for multilingual communication and accessibility. Translating content accurately from English to Arabic opens up opportunities to reach broader audiences, support global applications, and enable seamless interactions across regions where Arabic is the primary language. By leveraging the combined strengths of both the primary and secondary models, we aim to create a translation system that offers high accuracy, context-aware translation, and efficient performance. This ensures that the translator maintains linguistic nuances, delivers precise context adaptation, and is scalable across different platforms and applications.