Yardstick Task 3

For fine-tuning an AI model, the quality of the dataset is key. To ensure this, it’s important to focus on a few areas:

First, **data collection** needs to be aligned with the task. In my case, the dataset should revolve around questions and answers about software solutions and services. The more diverse the dataset (in terms of how users may phrase questions), the better the model will perform. It's also important to ensure that the answers are accurate and relevant, which might involve manual review or expert validation.

Once the data is collected, I would focus on **data augmentation**. This means creating variations of the questions in the dataset to make it more robust. For example, I might use paraphrasing techniques to generate different ways of asking the same question. This increases the model’s ability to understand and respond to various queries, improving its overall performance.

Data preprocessing also plays a big role. I would ensure that the text is clean, properly tokenized, and normalized so the model can better understand the input. Filtering out irrelevant or redundant data is crucial to keep the training process efficient.

Now, when it comes to fine-tuning the model itself, there are a few approaches. I would personally go with **supervised fine-tuning**. This method involves training the model on labeled data, where each input (like a question) has a corresponding label (the correct answer). Since I have access to a specific set of labeled data for Yardstick Software Solutions, this approach is ideal. It ensures that the model learns exactly what I need it to, without needing massive amounts of data or complex techniques like active learning. Plus, it gives me direct control over the training process, making it easier to ensure the model’s performance is aligned with my goals.

To sum it up, for this project, I’m focusing on collecting a strong, domain-specific dataset, augmenting it to ensure variety, and then applying supervised fine-

tuning. This method will help me create a model that answers questions accurately and understands the context behind them.