Dataset Preparation for Fine-Tuning

Techniques for Developing and Refining Datasets to Ensure High Quality for Fine-Tuning an Al Model:

a. Data Augmentation:

- Paraphrasing customer issues and tech responses to create variations and help the model generalize better to diverse phrasing.
- Synthetic Query Generation: You could generate new customer issues using Al models like GPT or fine-tune a paraphrasing model to generate additional examples of common support questions or issues.
- This will ensure that the fine-tuned model is exposed to varied input and is more robust in handling different user queries.

b. Quality Control and Annotation:

- Since the dataset is centered around customer issues and technical responses, it's crucial to ensure that all data is well-annotated with accurate labels. For instance.
- Use expert annotators to label Tech Response entries as correct, partially correct, or needs revision, ensuring that the model learns to prioritize responses that provide the most accurate help.
- Implement automated tools or human checks to filter out any low-quality responses (e.g incomplete answers) to improve the dataset's overall quality.

c. Dataset Filtering:

Filtering noisy data is essential. Filtering out any queries or responses that:

- Do not make sense in the context of the technical support task.
- Contain errors or irrelevant content, such as responses that don't properly address customer issues.
- Also, ensuring that the dataset represents all types of customer issues (e.g., technical, account related, or usage-related issues) will prevent bias and help the model respond appropriately to a wide range of queries.

d. Domain-Specific Data:

For fine-tuning the generative model (like Gemini Al), collected more domain-specific technical support data. If the dataset is broad but not specific to particular tech support scenarios, augment it with datasets that focus on very specific troubleshooting steps, FAQs, or common technical problems and solutions.

This will ensure that the model performs better in a highly specialized domain and can generate more relevant and contextually urate responses.