AI Engineer Assignment – Customer Support Questions Clustering

Objective:

Create a pipeline that will group customer questions into meaningful clusters and return the results in structured JSON format.

Instructions:

1. **Dataset**

   Use the customer questions dataset from this GitHub repo:

   🔗 https://github.com/bitext/customer-support-llm-chatbot-training-dataset/tree/main

2. **Task**

   - Construct a pipeline that classifies the questions into meaningful clusters.
   - Do **not** expose or use the original labels from the dataset. Use only the raw questions as input to the model.

3. **Output**

   For each identified cluster, generate a structured JSON with the following fields:

   - `"name"`: A short title summarizing the group of questions.
   - `"description"`: A clear explanation of what kinds of questions belong to this cluster.

- `"count"`: Number of questions that fall into this cluster.

Sample Output Format:

```
[
  {
    "name": "Pricing Inquiries",
    "description": "Questions related to subscription plans, costs, discounts, or billing issues.",
    "count": 42
  },
  {
    "name": "Technical Issues",
    "description": "Questions about bugs, errors, or problems using the product features.",
    "count": 31
  }
]
```

4. **Accuracy Estimation**

   After clustering, compare your extracted clusters with the **original dataset labels** to estimate clustering accuracy.

Deliverables

- Python code in Github implementing the pipeline
- Final JSON output
- Evaluation report