Q1.Why was the model/architecture used?

Ans: I used Chat-Gpt4o, because it is efficient and low in size and provide comparable accuracy as ChatGpt-4

Q2. What were the different factors considered for generating this dataset? (Length, topic Diversity etc.)  
  
Ans: **Length**: Ensuring reviews mimic real-world Amazon reviews with short, medium, and long text variations.

**Topic Diversity**: Include reviews on various supplements, covering multiple aspects such as ingredients, efficacy, taste, side effects, and packaging.

**Language Complexity**: Use a range of language styles, from simple and straightforward to more complex reviews, imitating both technical jargon and layman's terms.

**Sentiment**: Maintain a balance of positive, negative, and neutral reviews to capture the emotional tone seen in real reviews.

Q3. How do we measure the efficacy of a synthetic dataset?  
  
Ans: You can use the Predefined Metrics like: Rougue, pre-process the dataset, use the model which is good for synthetic data generator also use the synthetic dataset to train models for a given task (e.g., sentiment analysis or topic modelling) and compare the model performance to that achieved with real data.

Q4. How do we ensure the synthetic dataset one generates is inspired from a source dataset but not an exact replica?  
  
Ans**: Modelling Variability**: Implement randomness in the generation process by introducing slight variations in sentiment, length, or structure, ensuring the new dataset doesn’t copy exact reviews.

**Manual/Automated Checks**: Perform a manual or automated comparison (using cosine similarity or plagiarism detection tools) to ensure that the synthetic reviews are not direct copies of any reviews from the original dataset.

Q5. What were the top challenges in solving for this problem statement?

Ans **Realism vs. Variability, Evaluation the model, Pre-process the data set**